A categorical structure–activity relationship analysis of GPR119 ligands

Pritesh Kumar[a], Carl A. Carrasquer[bc], Arren Carter[cd], Zhao-Hui
Song[a] & Albert R. Cunningham[abc]

[a] Department of Pharmacology and Toxicology, University of
Louisville, Louisville, KY, USA

[b] Department of Medicine, University of Louisville, Louisville, KY,
USA

[c] James Graham Brown Cancer Center, University of Louisville,
Louisville, KY, USA

[d] Department of Chemistry University of Louisville, Louisville, KY,
USA
Published online: 17 Nov 2014.

PLEASE SCROLL DOWN FOR ARTICLE

Taylor & Francis
Taylor & Francis Group

A categorical structure–activity relationship analysis of GPR119 ligands

Pritesh Kumar[a], Carl A. Carrasquer[b,c], Arren Carter[c,d], Zhao-Hui Song[a]* and Albert R. Cunningham[a,b,c]*

[a]Department of Pharmacology and Toxicology, University of Louisville, Louisville, KY, USA; [b]Department of Medicine, University of Louisville, Louisville, KY, USA; [c]James Graham Brown Cancer Center, University of Louisville, Louisville, KY, USA; [d]Department of Chemistry University of Louisville, Louisville, KY, USA

The categorical structure–activity relationship (cat-SAR) expert system has been successfully used in the analysis of chemical compounds that cause toxicity. Herein we describe the use of this fragment-based approach to model ligands for the G protein-coupled receptor 119 (GPR119). Using compounds that are known GPR119 agonists and compounds that we have confirmed experimentally that are not GPR119 agonists, four distinct cat-SAR models were developed. Using a leave-one-out validation routine, the best GPR119 model had an overall concordance of 99%, a sensitivity of 99%, and a specificity of 100%. Our findings from the in-depth fragment analysis of several known GPR119 agonists were consistent with previously reported GPR119 structure–activity relationship (SAR) analyses. Overall, while our results indicate that we have developed a highly predictive cat-SAR model that can be potentially used to rapidly screen for prospective GPR119 ligands, the applicability domain must be taken into consideration. Moreover, our study demonstrates for the first time that the cat-SAR expert system can be used to model G protein-coupled receptor ligands, many of which are important therapeutic agents.

Keywords: G protein-coupled receptor 119 (GPR119); structure–activity relationship (SAR); diabetes; fragment-based modelling; categorical structure–activity relationship (cat-SAR)

1. Introduction

G protein-coupled receptors (GPCRs) are a large family of seven-transmembrane domain receptors that respond to diverse external signals and transmit information to signalling pathways inside the cell. GPCR activation via ligand binding often results in the generation of second messengers that regulate a broad range of physiological functions. G protein-coupled receptor 119 (GPR119) is a member of the class A (rhodopsin-type) GPCR family, which is highly expressed in pancreatic β-cells and in enteroendocrine cells of the gastrointestinal tract [1–4]. GPR119 agonists have been shown to increase insulin secretion and inhibit appetite [1–4]. As a result, GPR119 has recently emerged as a novel and promising therapeutic target for both type 2 diabetes (T2D) and obesity [1–4].

Agonist binding to GPR119, which is coupled to Gαs, a heterotrimeric G protein, results in an increase in intracellular cyclic adenosine monophosphate (cAMP) by activating

*Corresponding authors. Email: zhsong@louisville.edu (Z.-H. Song); al.cunningham@louisville.edu (A.R. Cunningham)

adenylate cyclase [5]. There are at least two mechanisms by which GPR119 agonists stimulate insulin release: 1) increased cAMP signalling results directly in an enhancement of glucose-dependent insulin release [6,7], or 2) increased glucagon-like peptide 1 (GLP-1) levels which further stimulates glucose-dependent insulin secretion [6,7]. Furthermore, GLP-1 inhibits glucagon secretion, appetite, and gastric emptying [7,8].

The broad interest in discovering novel, orally effective GPR119 agonists as potential therapies for T2D and obesity has resulted in the development of many synthetic GPR119 agonists over the past several years [1–4]. Recently, GPR119 agonists have already reached the stage of being investigated for clinical use [9–12].

Structure–activity relationship (SAR) modelling is a method designed to ascertain relationships between chemical structure and qualitative biological activity of ligands. Quantitative SAR and qualitative SARs are relationships that are derived from continuous data (e.g. biological potency) and non-continuous data (e.g. active or inactive), respectively.

The lack of an X-ray crystal structure for GPR119 hinders the ability to understand how ligands bind and interact with this receptor. As such, the aim of the current study was to vigorously investigate the relationship between chemical structure and biological activity by employing a fragment-based qualitative SAR expert system, cat-SAR, to study GPR119 ligand characteristics.

The cat-SAR expert system is flexible and user-friendly in the development of the learning set and model parameterization [13]. Cat-SAR analysis permits the user to designate adjustable modelling parameters including the selection of the size of the two-dimensional fragments, inclusion or exclusion of hydrogen atoms in the analysis, and rules for selecting important fragments for the final model [13]. Hence, the selection of compounds included in the learning set and control over various model parameters provides the user with the ability to more thoroughly investigate the relationship between chemical structure and biological activity [13].

Cat-SAR models are built through a comparison of structural features found amongst categorized compounds (active and inactive) in the model's learning set [13]. Fundamentally, the cat-SAR approach is transparent in the development of the learning set, the identification of fragments, and the determination of important fragments [13]. Moreover, the approach permits a high degree of user involvement and model optimization during the modelling process. This method includes the ability to examine the entire fragment base, investigate and optimize the fragments that have hypothetical biological relevance. In previous analyses, the cat-SAR program was able to achieve an overall concordance between observed and predicted values of 92% for a set of chemicals assessed for their ability to induce respiratory hypersensitivity [13] and 78–84% for a set of rat mammary carcinogens [14].

Moreover, since cat-SAR is based on the analysis of categorical data and two-dimensional fragments versus intact chemicals, the program can examine data sets that are divided into categories of activity rather than degrees of potency as in the case of QSAR [13]. Thus, in contrast to Hansch and conformational molecular field analysis (CoMFA) approaches which require continuous-type data, cat-SAR functions by identifying molecular attributes associated with biological activity by comparing characteristics of active (e.g. compounds known to act as agonists for GPR119 to inactive (e.g. compounds known to not activate GPR119) compounds. The models and subsequent predictions based on this can be used to examine structural features associated with activity and predict the probability of activity of unknown compounds, respectively [13].

Recently, a hierarchical virtual screening study [15] was carried out to identify novel agonists for the β_3-adrenergic receptor, a GPCR. The approach consisted of pharmacophore

modelling, docking and virtual screening which resulted in the identification of possible leads as novel β_3-adrenergic receptor agonists [15].

Herein, we describe the development of several novel GPR119 SAR models using the cat-SAR expert system to analyse the structural attributes of compounds that activate GPR119 and report predictive and mechanistically insightful SAR models for GPR119 activation. Overall, the cat-SAR models discussed herein for GPR119 activation demonstrate a high degree of predictive ability and mechanistic interpretability and may be useful for screening new drug candidates for this GPCR. These models can potentially be used to virtually screen large compound libraries to identify novel GPR119 ligands.

2. Materials and methods

2.1 *Materials*

The cat-SAR models are generated through an evaluation of structural features found amongst two designated categories of compounds in the model's learning set: Active or Inactive. The cat-SAR learning set consists of the chemical name, its structure as a MOL2 file, and its categorical designation (e.g. one or zero for active and inactive, respectively). Typically, organic salts are included as the freebase and simple mixtures and technical grade preparations may be included as the active component. Metals, metalo-organic compounds, polymers, hydrogen atoms, and mixtures of unknown composition are excluded.

The active data set consisted of 222 compounds that were collected from literature sources (see Supplementary Material, available via the Supplementary Content tab on the article's online page at http://dx.doi.org/). The inactive data set (See Supplementary Material) consisted of compounds determined not to activate GPR119 (less than 10% of AR231453 activity) at a concentration of 10 μM. Previously, we have validated a high-throughput cAMP assay for screening GPR119 ligands [16]. Using this assay, in the current study, we experimentally tested the compounds in three commercially available libraries (FDA-approved drug library, NIH clinical collection, and Tocriscreen) as potential GPR119 agonists. The FDA-approved drug library was purchased from Enzo Life Sciences (Farmingdale, NY). The NIH clinical collection was purchased from Evotec, Inc. (San Francisco, CA). The Tocriscreen library was purchased from R and D Systems, Inc. (Minneapolis, MN).

Our experimental screen did not result in the identification of any agonists, but resulted in the determination of 1000 inactive compounds. Four sets of 222 randomly selected inactive compounds were produced to generate four replicate models (standard 222 active and random 222 inactive). Therefore, we were able to assess the stability of the derived models. This approach prevented the chance of selecting 222 inactive compounds that produced a 'good' model. Therefore, we built four models consisting of 222 active and random sets of 222 inactive compounds.

The cat-SAR program provides for a number of user-specified options, so there is no a priori determination of the parameters in the final model. As such, we have developed and reported herein four different cat-SAR GPR119 models. With the ability to vary modelling parameters some can extend past the structural range of the learning sets and must be taken into consideration. For example, the fragment length parameter for the models described herein was set from three to seven heavy atoms (described below). Thus, chemicals of only three heavy atoms contributed their entire chemical structure as one fragment. Likewise, compounds consisting of less than three heavy atoms contributed no fragments to the model.

2.2 *Methods*

2.1.1 *In silico chemical fragmentation and fragment clustering*

Previous cat-SAR models used the Tripos Sybyl HQSAR module to generate chemical fragments. We have developed a novel algorithm for the in silico fragmentation of compounds. For each compound the respective MOL2 file was used to generate a computational unordered graph, represented by $G(V,E)$ where V is the set of vertices (atoms) and E is the set of edges (bonds) that connect a given pair of vertices. Next, each vertex was iterated over and all unique, connected subgraphs within six edges – the maximum fragment length – containing that vertex were identified, after which the given root vertex was removed from the graph for the remaining iterations. These subgraphs serve as mathematical representations of the chemical fragments. To convert the subgraphs to usable canonical SMILES, a Depth First Search of each subgraph was performed and the resulting SMILES was assigned using methodology derived from the CANGEN process of Daylight Chemical Information Systems.

As in previous cat-SAR models [14,17,18], chemical fragments that serve as valuable descriptors of activity/inactivity were identified and retained. However, there remained a high degree of redundancy between many of these fragments (based on similar chemical structures and derivation from mostly the same compounds). To ease in model interpretation and increase model accuracy and efficiency, this redundant fragment information was condensed by clustering the fragments. The clustering methodology utilizes the Tanimoto Similarity Coefficient and compound derivation similarity to determine relatedness between any two fragments. If two fragments share a Tanimoto Coefficient ≥70% and are present in ≥70% of the same compounds, those two fragments are then determined to be related. Once every possible combination of two fragments in the model was tested for relatedness, a second graph was generated with the vertices representing fragments and the edges representing relationships (either related or non-related). A clustering algorithm was then used to generate all fragment clusters. The clusters contained anywhere from a single fragment to over a hundred fragments, with each clusters activity being representative of the activity of each of their members.

2.2.2 *Identifying 'important' fragment and fragment clusters of activity and inactivity*

As mentioned, four fragment models were developed leading to the ultimate development of one cluster model (our final model). These four fragment models were used for preliminary analysis and the best model was chosen for cluster analysis and final model (cluster model) development. The general mechanism for identifying and selecting fragments or fragment clusters are similar, and are described together.

To determine any association between each fragment or fragment cluster and biological activity (or inactivity), a set of rules was implemented to select 'important' active and inactive clusters. The first selection rule – or the Number Rule – is the number compounds in the learning set that contain fragment(s) derived from a given cluster, which – in this exercise – was set at between three and five compounds. Looking at clusters that come from between three and five compounds in the learning set, models derived in the three to five range would be more inclusive (i.e. higher coverage), while those in the four to five range would be more accurate (i.e. higher concordance).

The second rule concerns the proportion of active or inactive compounds that contribute to each cluster and in this study ranged from between 85% to 95%. Even if a particular cluster is associated with activity, there may be other factors (i.e. clusters) that contribute to it

being inactive, and would not be expected to be found in 100% of the active compounds. For inactive fragments, a comparable argument can be made. Thus, by taking into account clusters toward the lower high end of the proportion scale (e.g. derived from 60% active and 40% inactive) model would be expected to again be more inclusive (i.e. higher coverage) while those derived from the higher end of the proportion scale (e.g. 90% active and 10% inactive) would be more accurate (i.e. higher concordance).

2.2.3 *Rule optimization*

As in previous cat-SAR models [14,17,18], setting of parameters for selecting important fragments (fragment compound counts and fragment activity proportion values) was used, with this experiment applying the same rules to fragment clusters. For these analyses, a rule optimization routine was employed wherein the Number Rule varied between 1 and 9 fragments or fragment clusters and the Proportion Rule varied between 0.50 and 0.95. Leave-one-out (LOO) validations were then performed for each model. The final models were chosen that were both highly accurate (i.e. had a high concordance between experimental and predicted values) and highly inclusive (i.e. made predictions on >90% of the chemicals in the learning set).

2.2.4 *Model validation*

A self-fit (i.e. leave-none-out (LNO)) and two cross-validations (i.e. LOO and multiple leave-many-out (LMO)) were performed for each model. The purpose of the self-fit analysis was to determine if the model that was built could be used to predict the activity of the chemicals in its learning set to confirm that the model could at least fit its own data [18] as well as mechanistic studies since all available data is used to generate a final model.

For the LOO cross-validation, each chemical, one at a time, was removed from the total fragment or cluster set and the $n - 1$ model was derived. For the LMO cross-validation, randomly selected sets of 10% of the chemicals were removed from the total cluster set and the $n - 10\%$ model was derived. Using the same criteria described above, the activity of the removed chemical was then predicted using the $n - 1$ model or $n - 10\%$ model. Predicted versus experimental values for each chemical or for the chemicals in the left-out sets were then compared and the model's concordance, sensitivity, and specificity were calculated [18], where Concordance = Correct predictions / Total predictions, Sensitivity = Correct positive predictions / Total positive predictions, Specificity = Correct negative predictions / Total negative predictions.

The cat-SAR predictions are based on the active and inactive fragment clusters. The predicted activity of a chemical is calculated based on the average probability of all the active and inactive compounds contributing to its fragment clusters. One method to classify compounds back to an active or inactive category is to determine an optimal cut-off point that best separates the probabilistic prediction of active and inactive compounds derived from the LOO validations [17]. Depending on the purpose of the model, the cut-off point can be adjusted wherein a model with the best overall concordance can be selected (i.e. a most predictive model), one with equal sensitivity and specificity (i.e. a balanced model that does not overly predictive active compounds at the cost of wrongly predicting inactive ones and vice versa), or one with high sensitivity.

2.2.5 *Predicting activity*

The resulting list of fragment clusters can then be used for mechanistic analysis, or to predict the activity of an unknown compound from the final model [17,18]. In order to predict the activity of an unknown compound, the cat-SAR program determines which, if any, clusters from the model's collection of important fragment clusters are present in the unknown or test compound [17,18]. If none are present, no prediction of activity can be made for the compound (i.e. there are no default predictions of inactivity or activity). If one or more clusters are present, the number of active and inactive compounds containing each cluster is determined and the probability of activity or inactivity is then calculated based on the total number of active and inactive compounds that went into deriving each of the fragment clusters [17,18].

The probability of activity was calculated with the cat-SAR FragSum routine [17,18]. This method calculates the average probability of the active and inactive clusters contained in each compound and is weighted to the number of active and inactive compounds that contribute to each cluster. For example, if a compound contains two clusters, one being found in 9/10 active compounds in the learning set (i.e. 90% active) and the other being found in 3/3 inactive compounds (i.e. 0% active), the unknown compound will be predicted to have a probability of activity of 69% (i.e. 9/10 actives + 0/3 actives = 9/13 actives or 69% chance of activity).

3. Results and discussion

3.1 *Overview of predictive performance of the cat-SAR GPR119 models*

The self-fit analysis of all models yielded concordance between experimental and predicted results averaging 99%. Considering the LOO validations with the FragSum method to calculate the probabilities of activity, the best GPR119 model had a concordance of 99%, a sensitivity of 99%, and a specificity of 100% (Model 1, Table 1). This model made predictions on 438 of the 440 chemicals in the learning set (no default prediction, see section 2.6 for description). The GPR119 models were also cross-validated with LMO. The GPR119 Model 1 had a concordance of 97%, a sensitivity of 95%, and a specificity of 99%, (Table 1).

3.2 *Comparison of models*

Using the difference between two proportions test, analysis of each set of four models derived from the random selection of inactive compounds indicated that the models had approximately the same concordance. For example, there was no significant difference between the four models. Model 1 correctly predicted 438 correct compounds out of 439 predictions (99%), and Model 2 correctly predicted 434 compounds out of 435 predictions (99%) ($p = 0.02$). Likewise, Model 3 correctly predicted 437 compounds out of 439 predictions (99%), and Model 4 correctly predicted 434 compounds out of 435 predictions (99%) ($p = 0.01$). This indicates that the accurate predictions made by the models were not spurious events based on a fortuitous random selection of 'good' compounds (i.e. random selections of 222 inactive compounds from the 1000 compound inactive set) and thus provides assurance that the models are based on a sound foundation and are not providing arbitrary predictions or mechanistic assertions. Since there was no significant difference between the four fragment models, we used Model 1 for cluster analysis and final model (cluster model) development.

Table 1. Fragment summary, self-fit, and cross-validation results for the GPR119 models.

Fragments	Model 1	Model 2	Model 3	Model 4	Clustered Model 1
Total	19028	18828	18966	18313	554
Model	5635	4036	8756	5257	121
Active	2621	2015	3420	2631	64
Inactive	3014	2021	5336	2626	57
Self-fit					
Sensitivity	99% (216/217)	99% (212/213)	99% (215/217)	99% (212/213)	98% (211/216)
Specificity	100% (222/222)	100% (222/222)	100% (222/222)	100% (222/222)	99% (219/221)
Concordance	99% (438/439)	99% (434/435)	99% (437/439)	99% (434/435)	98% (430/437)
Leave-one-out					
Sensitivity	96% (210/219)	96% (207/216)	96% (208/218)	95% (206/216)	99% (201/204)
Specificity	96% (213/221)	96% (212/221)	96% (212/222)	96% (212/222)	99% (192/194)
Concordance	96% (423/440)	96% (419/437)	96% (421/440)	95% (418/438)	99% (393/398)
Leave-many-out					
Sensitivity	95% (19.6/20.7)	92% (18.8/20.4)	92% (19.2/20.7)	95% (19.4/20.5)	83% (17.1/20.5)
Specificity	99% (20.5/20.8)	99% (20.8/20.9)	99% (20.7/20.9)	99% (20.6/20.9)	99% (20.5/20.7)
Concordance	97% (40.1/41.5)	96% (39.6/41.3)	96% (39.8/41.6)	97% (40.1/41.4)	91% (37.6/41.3)

3.3 *Analysis of compounds in the training set*

The two significant pharmacophores consist of: (1) an aryl or heteroaryl moiety substituted with a hydrogen bond-accepting group on one part of the molecule and (2) a piperidine moiety N-capped with a carbamate or an isosteric heterocycle on the opposite side of the molecule. These two motifs are connected via an appropriate central spacer containing a heterocyclic ring or an acyclic chain [19].

Our clustering analysis of AR231453 indicates that the nitro-pyrimidine core and the presence of a sulfone moiety are responsible for AR231453 agonist activity. Furthermore, analysis of AR231453 resulted in several clusters that indicated the presence of critical hydrogen bond acceptors that is consistent with a previous report [20]. In addition, several clusters contained a sulfone group that has previously been described as a key functional group for AR231453 as shown in Figure 1. Specifically, our clustering analysis of AR231453 resulted in the generation of five clusters (Cluster 53, Cluster 113, Cluster 129, Cluster 177, and Cluster 503), of which a representative fragment from each group is shown (Figure 1), which are associated with the activity of this compound.

The SAR analysis of PSN632408 has also been described [19]. Through extensive SAR analysis, it was determined that the N-capped piperidine motif is required for GPR119 agonist activity. Our clustering analysis of PSN632408 resulted in the generation of five clusters (Cluster 54, Cluster 135, Cluster 188, Cluster 234, and Cluster 444), of which a representative fragment from each group is shown, that are associated with the activity of this compound (Figure 2). This analysis is consistent with a previous report [19], wherein the
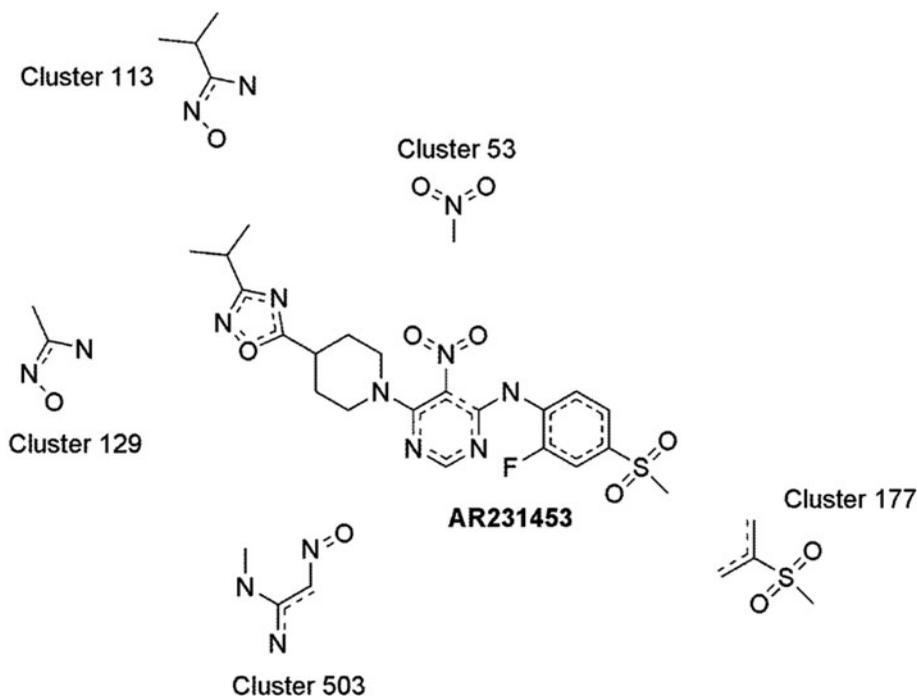

Figure 1. Five representative clusters used to predict the activity of AR231453 by the GPR119 cat-SAR model.
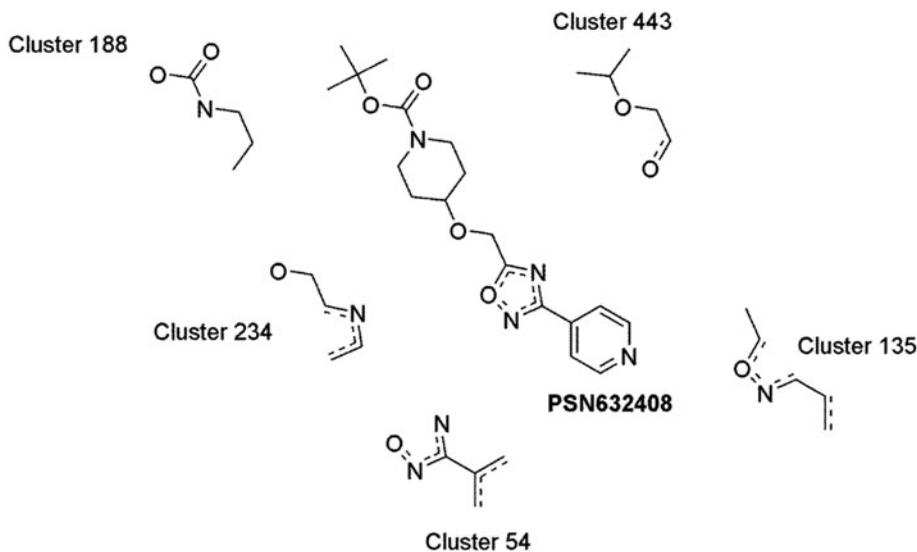
Figure 2. Five representative clusters used to predict the activity of PSN632408 by the GPR119 cat-SAR model.

N-capped piperidine core with a carbamate was determined to be responsible for PSN632408 agonist activity.

3.4 *Analysis of compounds not in the training set*

Several of the known GPR119 agonists have evolved from the prototypical compounds 2-fluoro-4-methanesulfonyl-phenyl)-{6-(4-(3-isopropyl-(1,2,4)oxadiazol-5-yl)-piperidin-1-yl}-5-nitro-pyrimidin-4-yl}-amine (AR231453) [20] and tert-butyl 4-{(3-pyridin-4-yl-1,2,4-oxa-diazol-5-yl)methoxy}piperidine-1-carboxylate (PSN632408) [21].

In addition to the internal validations performed (LOO, LMO), we performed an external validation in which we used our model to predict the activities of 45 compounds (external test set) which were not present in the training set. The purpose of this was to ensure the robustness of our model by testing the hypothesis that it could accurately predict the activity of compounds not in our training set.

Of these 45 compounds, 14 were known GPR119 agonists while 31 were not agonists for GPR119 confirmed by cAMP assay. The 14 known agonists consisted of compounds that can be structurally classified in two groups: 1) bicyclic amine scaffolds and 2) pyrazolopyrimidine scaffolds. These 14 known GPR119 agonists were selected from patents and literature sources (see Supplementary Material). The 31 inactive compounds consisted of compounds which do act as agonists for GPR119 at concentrations up to 10 μM, which we experimentally confirmed using our established HTRF assay [16]. These inactive chemicals consisted of compounds originating from three distinct chemical libraries (FDA, NIH clinical collection, Tocriscreen) ensuring structural diversity (see Supplementary Material). Our GPR119 model 1 had a 78.6% success rate in correctly predicting 11 out of 14 as GPR119 agonists. Furthermore, our model had a 90.3% success rate in correctly predicting 28/31 as inactive

compounds (not GPR119 agonists). Overall, the model achieved an 86.7% success rate in correctly predicting the activity of compounds in the test set.

Recently, Wu et al. synthesized a series of piperazinylpyridine derivatives as GPR119 agonists [22]. Through SAR analysis, compounds with alkylsulfonamide and isopropylcarbamate end groups displayed potent GPR119 receptor activity [22]. Our clustering analysis of propan-2-yl 4-([6-[4-(propane-1-sulfonyl)piperzin-1-yl]-oxy)methylpiperidine-1-carboxylate (Wu Compound 19A) resulted in the generation of three clusters (Cluster 36, Cluster 64, and Cluster 435), of which a representative fragment from each group is shown, which are responsible for the activity of this compound (Figure 3). Our fragment analysis indicates that the sulfonamide and carbamate groups are important for Wu Compound 19A agonist activity, which is consistent with previously reported SAR of this compound [22].

Previously, in a separate study, Sakairi et al. disclosed a novel series of GPR119 agonists based on a bicyclic amine scaffold [23]. Through SAR analysis of Wu Compound 19A, it was determined that the basic nitrogen atom of the bicyclic amine played an important role in the production of GPR119 agonist activity [23]. Furthermore, Sakairi and coworkers showed that the carbonyl group on the bicyclic core represented a better pharmacophore than a sulfonyl group [23], which is consistent with our results.

The first ligand-based pharmacophore model of GPR119 was developed by Zhu and coworkers to obtain a hypothetical picture of the chemical features responsible for activity [24]. Pharmacophore models were generated with 24 known GPR119 agonists using Discovery Studio V2.1 [24]. The application of this model was able to predict the activity of 25 known GPR119 agonists with a correlation coefficient of 0.933.

Wellenzohn and coworkers recently described the application of a virtual screening technique that was used to identify novel GPR119 agonists [25]. The virtual screening process consisted of an activity anchor and the use of feature tree fragment space searches which was followed by a 3D post-processing step [25]. The in silico results were then filtered and prioritized and combinatorial libraries of target molecules were synthesized. This method resulted in the discovery of two new structural classes of potent GPR119 agonists, one of which has progressed as a novel lead class [23].
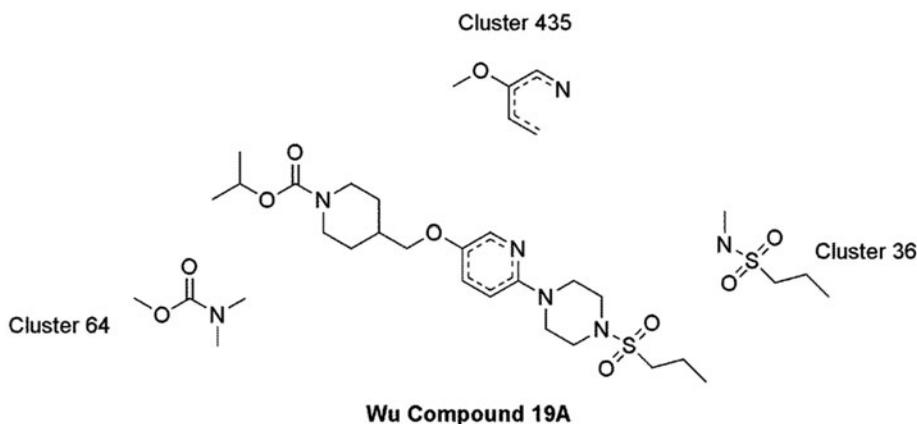

Figure 3.   Three representative clusters used to predict the activity of Wu Compound 19A by the GPR119 cat-SAR model.

A key difference between our modelling approach and Zhu and coworkers [24] is that cat-SAR is based on the analysis of categorical data and two-dimensional fragments versus intact chemicals [17]. This allows the program to examine data sets that are divided into categories of activity rather than degrees of potency [17].

3.5 *Applicability domain of models*

It should be noted that even though our models have high specificity, high sensitivity, and high concordance values, the predictive ability is limited by the model's applicability domain (AD). The AD refers to a theoretical region in the space defined by the descriptors of the model which provides insight into the development and applicability on which the training set can make reliable prediction for unknown compounds. As far as a potential virtual screening tool, the AD of our GPR119 models is somewhat constrained by the lack of diversity of the structures of the active set. To be useful as a virtual screening tool, it would be necessary to sort and rank the compounds that were predicted as potential ligands and compare the structures to the active compounds in the training set. Then, we could make ad-hoc decisions on whether or not to test these potential ligands.

4. Conclusions

We report, for the first time, the application of a fragment-based modelling approach using the cat-SAR expert system to model GPR119. The good predictive ability of these models to understand 2-D molecular fragments indicates their potential usefulness in investigating the relationship between GPR119 ligand structure and activity, as our model was able to correctly predict the activity of compounds outside of the training set.

It is expected that the generated information could be used to identify the chemical moieties specific to GPR119 activity. Thus the cat-SAR expert system produces models which are predictive and are based on mechanically sound attributes. Most importantly, our study is the first to demonstrate that the cat-SAR expert system can be used to model a GPCR. Overall, we have generated a model that can be potentially used to virtually screen large databases with high specificity, high sensitivity, and high concordance, but the AD must be taken into consideration.

Acknowledgments

References

[1] N.G. Morgan and S. Dhayal, *The significance of GPR119 agonists as a future treatment for type 2 diabetes*, Drug News Perspect. 23 (2010), pp. 418–424.

[2] H.A. Overton, M. Fyfe, and C. Reynet, *GPR119, a novel G protein-coupled receptor target for the treatment of type 2 diabetes and obesity*, Br. J. Pharmacol. 153 Suppl 1 (2008), pp. S76–81.

[3] R.M. Jones, J.N. Leonard, D.J. Buzard, and J. Lehmann, *GPR119 agonists for the treatment of type 2 diabetes*, Expert Opin. Ther. Pat. 19 (2009), pp. 1339–1359.

[4] U. Shah and T.J. Kowalski, *GPR119 agonists for the potential treatment of type 2 diabetes and related metabolic disorders*, Vitam. Horm. 84 (2010), pp. 415–448.

[5] Z. Chu, R.M. Jones, H. He, C. Carroll, V. Gutierrez, A. Lucman, M. Moloney, H. Gao, H. Mondala, D. Bagnol, D. Unett, Y. Liang, K. Demarest, G. Semple, D.P. Behan, and J. Leonard, *A role for beta-cell-expressed G protein-coupled receptor 119 in glycemic control by enhancing glucose-dependent insulin release*, Endocrinology 148 (2007), pp. 2601–2609.

[6] G. Flock, D. Holland, Y. Seino, and D.J. Drucker, *GPR119 regulates murine glucose homeostasis through incretin receptor-dependent and independent mechanisms*, Endocrinology 152 (2011), pp. 374–383.

[7] L. Lauffer, R. Iakoubov, and P.L. Brubaker, *GPR119: "Double-dipping" for better glycemic control*, Endocrinology 149 (2007), pp. 2035–2037.

[8] L. Baggio and D.J. Drucker, *Biology of incretins: GLP-1 and GIP*, Gastroenterology 132 (2007), pp. 2131–2157.

[9] J.W. Polli, E. Hussey, M. Bush, G. Generaux, G. Smith, D. Collins, S. McMullen, N. Turner, and D.J. Nunez, *Evaluation of drug interactions of GSK1292263 (a GPR119 agonist) with statins: From in vitro data to clinical study design*, Xenobiotica 43 (2013), pp. 498–508.

[10] L.B. Katz, J.J. Gambale, P.L. Rothenberg, S.R. Vanapalli, N. Vaccaro, L. Xi, T.C. Sarich, and P.P. Stein, *Effects of JNJ-38431055, a novel GPR119 receptor agonist, in randomized, double-blind, placebo-controlled studies in subjects with type 2 diabetes*, Diabetes Obs. Metab. 14 (2012), pp. 709–716.

[11] D.J. Nunez, M.A. Bush, D.A. Collins, S.L. McMullen, D. Gillmor, G. Apseloff, G. Atiee, L. Corsino, L. Morrow, and P.L. Feldman, *Gut hormone pharmacology of a novel GPR119 agonist (GSK1292263), metformin, and sitagliptin in type 2 diabetes mellitus: Results from two randomized studies*, PloS one 9 (2014), pp. e92494.

[12] L.B. Katz, J.J. Gambale, P.L. Rothenberg, S.R. Vanapalli, N. Vaccaro, L. Xi, D.C. Polidori, E. Vets, T.C. Sarich, and P.P. Stein, *Pharmacokinetics, pharmacodynamics, safety, and tolerability of JNJ-38431055, a novel GPR119 receptor agonist and potential antidiabetes agent, in healthy male subjects*, Clin. Pharmacol. Therap. 90 (2011), pp. 685–692.

[13] A.R. Cunningham, S.L. Cunningham, D.M. Consoer, S.T. Moss, and M.H. Karol, *Development of an information-intensive structure-activity relationship model and its application to human respiratory chemical sensitizers*, SAR QSAR Environ. Res. 16 (2005), pp. 273–285.

[14] A.R. Cunningham, S.T. Moss, S.A. Iype, G. Qian, S. Qamar, and S.L. Cunningham, *Structure-activity relationship analysis of rat mammary carcinogens*, Chem. Res. Toxicol. 21 (2008), pp. 1970–1982.

[15] A.K. Saxena and K.K. Roy, *Hierarchical virtual screening: Identification of potential high-affinity and selective beta(3)-adrenergic receptor agonists*, SAR QSAR Environ. Res. 23 (2012), pp. 389–407.

[16] P. Kumar, A. Kumar, and Z.H. Song, *Structure-activity relationships of fatty acid amide ligands in activating and desensitizing G protein-coupled receptor 119*, Eur. J. Pharmacol. 723 (2014), pp. 465–472.

[17] A.R. Cunningham, C.A. Carrasquer, and D.R. Mattison, *A categorical structure-activity relation-ship analysis of the developmental toxicity of antithyroid drugs*, Int. J. Pediatr. Endocrinol. 2009 (2009), 936154.

[18] S. Qamar, C.A. Carrasquer, S.L. Cunningham, and A.R. Cunningham, *Anticancer SAR models for MCF-7 and MDA-MB-231 breast cell lines*, Anticancer Res. 31 (2011), pp. 247–522.

[19] R.M. Jones and J.N. Leonard, *The emergence of GPR119 agonists as anti-diabetic agents*, Ann. Rep. Med. Chem. 44 (2009), pp. 149–170.

[20] G. Semple, B. Fioravanti, G. Pereira, I. Calderon, J. Uy, K. Choi, Y. Xiong, A. Ren, M. Morgan, V. Dave, W. Thomsen, D.J. Unett, C. Xing, S. Bossie, C. Carroll, Z. Chu, A.J. Grottick, E.K. Hauser, J. Leonard and R.M. Jones, *Discovery of the first potent and orally efficacious agonist of the orphan G-protein coupled receptor 119*, J. Med. Chem. 51 (2008), pp. 5172–5175.

[21] H.A. Overton, A.J. Babbs, S.M. Doel, M.C. Fyfe, L.S. Gardner, G. Griffin, H.C. Jackson, M.J. Procter, C.M. Rasamison, M. Tang-Christensen, P.S. Widdowson, G.M. Williams, and C. Reynet,

Deorphanization of a G protein-coupled receptor for oleoylethanolamide and its use in the discovery of small-molecule hypophagic agents, Cell. Metab. 3 (2006), pp. 167–175.

[22] Y. Wu, J.D. Kuntz, A.J. Carpenter, J. Fang, H.R. Sauls, D.J. Gomez, C. Ammala, Y. Xu, S. Hart, and S. Tadepalli, *2,5-Disubstituted pyridines as potent GPR119 agonists*, Bioorg. Med. Chem. Lett. 20 (2010), pp. 2577–2581.

[23] M. Sakairi, M. Kogami, M. Torii, H. Kataoka, H. Fujieda, M. Makino, D. Kataoka, R. Okamoto, T. Miyazawa, M. Okabe, M. Inoue, N. Takahashi, S. Harada, and N. Watanabe, *Synthesis and SAR studies of bicyclic amine series GPR119 agonists*, Bioorg. Med. Chem. Lett. 22 (2012), pp. 5123–5128.

[24] X. Zhu, D. Huang, X. Lan, C. Tang, Y. Zhu, J. Han, W. Huang, and H. Qian, *The first pharmacophore model for potent G protein-coupled receptor 119 agonist*, Eur. J. Med. Chem. 46 (2011), pp. 2901–2907.

[25] B. Wellenzohn, U. Lessel, A. Beller, T. Isambert, C. Hoenke, and B. Nosse, *Identification of new potent GPR119 agonists by combining virtual screening and combinatorial chemistry*, J. Med. Chem. 55 (2012), pp. 11031–11041.