Prediction of ligand-receptor pharmacological activities using a combined docking and machine learning approach.

Mireia Jiménez-Rosés[1,2,*], Bradley Angus Morgan[1,3,4], Maria Jimenez Sigstad[1,2], T.D. Zoe Tran[1,5], Rohini Srivastava[1,5], Asuman Bunsuz[1,2], Leire Borrega Román[1,2,6], Pattarin Hompluem[1,2], Sean A. Cullum[1,2,4], Clare R. Harwood[1,2,4], Eline J. Koers[1,2], David A. Sykes[1,2], Iain B. Styles[1,3,*] and Dmitry B. Veprintsev[1,2,*]

[1]Centre of Membrane Proteins and Receptors (COMPARE), University of Birmingham and University of Nottingham, Midlands, UK

[2]Division of Physiology, Pharmacology & Neuroscience, School of Life Sciences, University of Nottingham, Nottingham, NG7 2UH, UK

[3]School of Computer Science, University of Birmingham, Edgbaston, Birmingham, B15 2TT, UK and The Alan Turing Institute, London, NW1 2DB, UK

[4]MRC IMPACT Doctoral Training Programme, Universities of Birmingham, Leicester and Nottingham, Midlands, UK

[5]MSc Programme in Drug Discovery & Pharmaceutical Sciences, University of Nottingham, UK

[6]Department of Pharmacology, University of the Basque Country UPV/EHU, E-48940 Leioa, Spain

*correspondence should be addressed to Mireia.JimenezRoses@nottingham.ac.uk, I.B.Styles@bham.ac.uk and Dmitry.Veprintsev@nottingham.ac.uk.

Abstract

G protein coupled receptors (GPCRs) form one of the largest families of proteins in humans, and are valuable therapeutic targets for a variety of different diseases. One central question of drug discovery surrounding GPCRs is what determines the agonism or antagonism exhibited by ligands which bind these important targets. Ligands exert their action via the interactions they make in the ligand binding pocket. We hypothesised that there is a common set of receptor interactions made by ligands of diverse structures that mediate their action. We reasoned that among a large dataset of different ligands, the functionally important interactions will be over-represented. To investigate this hypothesis, we assembled a database of ~2700 known β_2AR ligands and computationally docked them to multiple experimentally determined β_2AR structures, generating ca 75,000 docking poses. For each docking pose, we predicted all interactions between the atoms of the receptor and the atoms of the ligand. Using Machine Learning (ML) we identified specific interactions that correlated with the agonist or antagonist activity of these ligands, and developed ML-based predictors of agonist/antagonist activity with up to 90% accuracy. This approach can be readily applied to other GPCRs and drug targets beyond GPCRs.

Introduction

G-protein-coupled receptors (GPCRs) remain a therapeutically important family of proteins with over 100 receptors targeted by 500 drugs approved for clinical use[1]. The human β_2-adrenoceptor (β_2AR) responds to stimulation by the endogenous agonist ligands adrenaline and noradrenaline by inducing Gs-mediated cAMP signalling and is a valuable target for small molecule smooth muscle relaxants used to treat asthma and other pulmonary diseases[2,3]. Endogenous agonist activity can be readily inhibited by so-called antagonist drugs that prevent receptor activation by occupying binding pocket without activation and blocking agonist access. All GPCRs share a common architecture of a bundle of seven transmembrane helices (TMs), with the ligand binding pocket accessible from the extracellular space and an intracellular effector binding site that becomes available following transition into an active receptor conformation[4]. One of the key features of GPCRs is that they are highly dynamic and adopt a number of distinct conformations that are important for engagement of signalling partners, eg. activation of the Gs protein or arrestins[5]. It is generally thought that ligands control GPCR activity by preferentially stabilising active or inactive conformations[6].

A large number of ligands have been developed to target β-adrenoceptors (βAR) over the last 60 years since the pioneering discovery of beta-blockers by Sir James Black[3,7-10]. β_2AR is one of the best studied GPCRs from a structural perspective, with 35 reported structures with 13 diverse ligands in inactive and active states reported so far. Numerous studies have attempted to link structural properties of the ligands to their functionality[11-14]. While these studies show significant promise in establishing the link between structure and function, the experimentally available structural data cover only a very small fraction of known β_2AR ligands which limits their general applicability. A key question in pharmacological drug research – what factors dictate ligand receptor agonism and antagonism – remains to this day open.

Ligands exert their action on GPCRs via the interactions they make in the ligand binding pocket. We hypothesised that despite the observed structural diversity of ligands targeting a particular receptor, there should be common interacting atoms within the ligand binding pocket that mediate their action. Given this hypothesis, we reasoned that among a large dataset of different ligands and their respective binding

poses, the functionally important atomic interactions the ligands make with a particular receptor will be over-represented. To investigate this hypothesis, we assembled a database of ~2700 known β_2AR ligands and computationally docked them to multiple experimentally determined β_2AR structures, generating ca 75,000 docking poses (Figure 1A and B). For each of the docking poses, we generated a detailed Atomic Interaction Fingerprint (AIF), which comprises of a list of all the pairs of atoms involved in interaction between a receptor and a ligand and the types of the bonds formed. In total, there were ca 1,100 possible interaction descriptors that we interchangeably call features (Figure 1C) in our dataset. Using pairwise correlation and Machine Learning (ML) approaches, we identified specific interactions between the ligands and the β_2AR that correlated with their reported agonist or antagonist activity at the receptor (Table S1). In addition to a common set of interactions that were present for both ligand types, agonists make specific contacts with the amino acid residues H93$^{2.64x63}$, K97$^{2.68x67}$, S203$^{5.42x43}$, S204$^{5.43x44}$, S207$^{5.46x461}$, H296$^{6.58x58}$ and K305$^{7.35x34}$ in transmembrane helices TM2, TM5, TM6, TM7 while antagonists make specific interactions with W286$^{6.48x48}$ and Y316$^{7.43x42}$ in TM6 and TM7. The initial ~2700 molecule data set allowed us to develop both pairwise correlation and ML-based predictions of ligand activity with a 70-90% success rate, validated via a PubChem metasearch that identified activity for an additional 349 ligands not included in the initial training set. This approach successfully identifies the key features of the ligands in terms of the individual interactions they make with the receptor to exert their pharmacological action. Moreover, it provides a novel strategy for understanding the molecular mechanism of drug action on receptors and represents a valuable approach for future rational drug design.

Results

β_2AR agonists are on average bigger and more lipophilic compared to antagonists.

To construct our dataset of currently known β_2AR ligands we searched all available open access repositories such as GPCRdb, ChEMBL, DrugBank, Guide to Pharmacology and ZINC. The curated database included 2683 unique β_2AR ligands, of which 1317 had reported pharmacological action (987 agonists and 330 antagonists/ inverse agonists). The remaining 1366 were classified as "known binders" with no assigned pharmacological activity (Figure 1A).

To understand if there are any obvious differences between agonists and antagonists, we compared their physicochemical (PC) properties predicted using OpenBabel software[15]. In general, PC property values for agonists were statistically different from those for antagonists (unpaired t-test, $p < 0.0001$), as seen in molecular weight (MW) and lipophilicity (logP) (Figure 2A and B, respectively). The MW of ~70% of agonist ligands was in the 350-550 g/mol range, with an average of 469 ± 108 g/mol. In contrast, the antagonist ligands were typically smaller, with ~70% within a range of 200-400 g/mol (average 358 ± 108 g/mol). The logP values of ~70% of agonists are in the range 3-7, with an average of 4.6 ± 1.6, whereas ~70% of antagonist ligands had logP values in the range 0-5 (average 3.1 ± 1.4). On the other hand, we observed an identical linear correlation between the molecular weight and lipophilicity for both agonists and antagonists (Figure 2C). Taken together, the β_2AR agonists profiled here tended to be more lipophilic and bigger in size but belong to the same class of compounds in respect to these two parameters.

Generating Atomic Interaction Fingerprints based on Molecular Docking Poses

To obtain structural information on how ligands in the curated dataset interact with the receptor (ie, ligand binding poses), we performed molecular docking using the open-source AutoDock Vina software[16]. Three β_2AR structures were studied: the active conformational states i) PDB 3SN6 stabilised by the Gs protein[17] and ii) PDB 4LDE stabilised by a nanobody[18], and iii) the inactive conformational state PDB 5JQH[19]. We obtained ~75,000 binding poses in total, ~25,000 poses for each PDB (up to 10 poses

for ligand, for 2683 compounds) (Figure 1B). Each ligand binding pose was used to generate an atomic interaction fingerprint (AIF) using Arpeggio software[20], in total we obtained ~75,000 AIF files (Figure 1C). Each AIF included ~60 unique interactions on average between the atoms of the ligand and atoms of the receptor. When the type of atoms of each ligand and the type of bond formed are taken into account, this resulted in over 1,100 possible types of interaction across the complete ligand dataset.

It is important to consider that the obtained AIF fingerprint dataset contains a lot of noise because not all obtained docking poses are likely to be relevant or functionally important. The limitations of the ligand docking algorithms result in multiple alternative binding poses with very similar "quality scores", with only one of the top ten solutions likely corresponding to the experimentally observed binding pose. While crystallographic structures typically represent one ligand binding pose, they tend to represent the lowest energy state of the system. Most often there is only one structure available due to the significant experimental effort needed to obtain a single GPCR structure. On the other hand, molecular dynamics simulation and biophysical experiments suggest that ligands are dynamic when bound to the receptor[21]. Therefore, it is important to consider multiple ligand docking poses in the analysis. We rationalised that in a large dataset of different ligands and their respective binding poses, the functionally important atomic interactions between the ligands and the receptor will be over-represented while the influence of the noise (irrelevant binding poses) would average out.

We improved the signal-to-noise ratio within our dataset by excluding irrelevant binding poses using prior knowledge based on crystallographic data (Figure 1C, filtering panel). The majority (ca 97%) of β_2AR ligands have a prevalent β-hydroxy-amine motif that makes specific interactions with the receptor. We therefore excluded poses that did not display this ionic interaction between the oxygen of D113$^{3.32x32}$ and the nitrogen atom of ethanolamine of the ligands and the hydrogen bond between the oxygen atom of N312$^{7.39x38}$ and either the NH or beta-hydroxyl groups in the ligand scaffold; these have been observed in every experimental crystallographic structure of the β_2AR. After applying this filter, we obtained ~31,500 atomic interactions files (~10,500 poses and AIF files for each PDB), reducing the size of the original dataset by ~55%. We refer to this as the "filtered dataset". As the filtering step also removed ~3% of ligands in our dataset that did not contain the β-hydroxy-amine motif or did not

produce suitable poses, we have also included in our analysis the “full dataset” consisting of~75,000 AIF files with no filtering for comparison.

Data-driven analysis reveals key interactions that drive agonism and antagonism of ligands

We constructed a ligand-receptor interaction matrix, organising the atom-atom interaction and type of interaction in the columns and each binding pose in rows for each PDB. We defined the ligand binding site as all residues that interact with at least one ligand binding pose in the dataset resulting in 30 residues in total (Table S2). The atoms of the ligand binding site provide a constant reference coordinate system to describe ligand-receptor interactions. We defined atomic interaction between specific atoms of the receptor, and the specific atom (C, N, O etc) in the ligand and the nature of the interacting bond (polar, ionic, hydrophobic, etc). This strategy allowed us to encode the ligand-receptor interaction matrix that accommodates diverse ligands irrespective of their structural scaffold.

Using Pearson’s pairwise correlation between the variables describing the presence or absence of an atomic interaction and the variable denoting agonist/antagonist properties of the ligands, we identified atom-atom interactions that associated with agonism or antagonism in the filtered dataset. From about 100 commonly observed interactions, we find that the most representative interactions for agonist ligands are hydrophobic/aromatic contacts involving K97$^{2.68x67}$, F194^{ECL2} , H296$^{6.58x58}$ and K305$^{7.35x34}$ and polar/ionic/hydrogen bond contacts with S203$^{5.42x43}$, S204$^{5.43x44}$, S207$^{5.46x641}$ and H296$^{6.58x58}$. The antagonists made specific hydrophobic/aromatic contacts with W286$^{6.48x48}$ and Y316$^{7.43x42}$ and polar/ionic/hydrogen bond contacts with Y316$^{7.43x42}$ (Figure 3A and Table S3).

While the majority of interactions had the same impact on receptor function (mediating agonism or antagonism) for all atoms of the individual residue, in some cases (D113$^{3.32x32}$, D192$^{45.51x51}$, F193$^{45.52x52}$, T195 ^{ECL2}, F289$^{6.51x51}$, F290$^{6.52x52}$, Y308$^{7.35x34}$, N312$^{7.39x38}$) this depended on the individual atoms of the residue and the nature of the interacting bond (Table S3). For example, the polar/ionic/hydrogen contact of the carbonyl oxygen (OD1, as defined by the Protein Data Bank format[22]) of D113$^{3.32x32}$ with an oxygen atom of a ligand is predictive of agonism while interaction with a

nitrogen atom is predictive of antagonism. Contacts made by hydroxyl oxygen (OD2) of D113$^{3.32x32}$ have the opposite effect: interaction with a nitrogen atom of the ligand corresponds to agonism, whilst interaction with an oxygen atom results in antagonism. In another example, polar contacts of the sidechain nitrogen (ND2) of N312$^{7.39x38}$ with oxygen atoms in the ligand corresponded to agonism whilst interaction with nitrogen leads to antagonism.

The full dataset was a more complex challenge as it contains more noise in terms of the number of different poses and also a more diverse range of ligands. Nonetheless we also observed around 100 common interactions, which were mostly the same as those determined for the filtered dataset. However, several interactions changed their relative importance (Figure 3B); for example, the importance of S204$^{5.43x44}$ as a determinant of agonism was reduced, while W313$^{7.40x39}$ became more predictive of agonism. However, the core set of agonist-associated interactions made with S203$^{5.42x43}$, S207$^{5.46x461}$ and, F194^{ECL2}, H296$^{6.58x58}$, K305$^{7.32x31}$ and K97$^{2.68x67}$ remained the same.

Having identified the key ligand-receptor atomic interactions (or features) that correlate with agonism or antagonism of the ligands, we used them to develop predictors of pharmacological activity based on the ligand binding pose and its AIF. Using the Pearson's pairwise correlation coefficients derived above, we calculated the cumulative score for all atomic interactions present, ranging from 0 for antagonists to 1 for agonist after normalization. Using a cut-off score of 0.37 (selected by maximising Matthews Correlation Coefficient, MCC) for the filtered dataset we obtained a pharmacological classification (agonist or antagonist) with a MCC of 0.43 that corresponds to the accuracy of prediction of 79% (Figure S1A). MCC is a measure of the quality of binary classifications when the classes are of different sizes as in our case (ca 75% are agonists). For the full dataset (cut-off = 0.51), the MCC and accuracy decreased to 0.29 and 67%, respectively (Figure S1B). An important consideration for interpretation of the prediction accuracy is that the training dataset may contain errors: compounds that are "wrongly" assigned to a particular class. Therefore, we would not expect the predictors to be 100% accurate during the validation step.

As the pairwise correlation approach identifies the relative importance of individual interactions, we applied ML strategies (see methods for details) that are capable of

detecting more complex patterns in the data than pairwise correlation analysis. We applied Random Forest Classifier (RFC)[23] on the filtered dataset and XGBoost[24] on the full dataset. RFC constructs a multitude of decision trees and averages them to improve the predictive performance and control overfitting, reaching MCC values in training of 0.81 and an accuracy of 92% on the filtered dataset (Figure S2 and S4A). The `XGBoost` algorithm that iteratively constructs optimised decision trees guided by the results of the previous steps performed remarkably well on the full dataset (Figure S3 and S4B), with a prediction performance on the holdout set of 0.78 MCC and 93% accuracy after full Bayesian optimisation. This suggests that there are predictive patterns in both the filtered and full dataset not captured by a simple predictor based on pairwise correlations.

It is, however, a considerable challenge to interpret what the `ML` algorithm has actually learned. We extracted the feature importance for RFC trained on the filtered dataset (Figure 4A,B) and the feature importance for XGBoost trained on the full dataset (Figure 4C,D), using the Shapley Additive Explanations (SHAP) values which reflect the contribution of each feature to the prediction. In most cases, the presence of a particular interaction is predictive of agonism or antagonism. However, in a minority of cases, the absence of the interaction was more important for predictions (eg, 193/CB-1/C hydrophobic).

Overall, while the relative order of importance of individual features varied depending on the model, we observed the same set of interactions that were predictive of agonism or antagonism for all two models (Table S4). The application of pairwise correlation analysis and ML methods allowed us to identify the key interactions associated with agonism or antagonism of ligands (Figures 3 and 5) and to develop effective predictors of their pharmacological activity based on their binding pose.

Validation of ligand pharmacological activity predictions

Our ligand dataset contained 1366 known binders for which the pharmacological activity was not reported in the source databases. We used the pairwise correlation and ML-based predictors to calculate their pharmacological activity based on their AIF, calculating the average for all the binding poses obtained (Table S5).

A posteriori, PubChem metasearch of 730 ligands from the “known binders” dataset belonging to three predicted classes: agonists, antagonists and ligands with intermediate predictions (53% of the whole binders set) allowed us to find reported pharmacological activity data for additional 399 compounds (331 agonists and 69 antagonist) to validate our predictions. As the drug discovery programs have been focused on developing agonists of β_2AR, antagonists are underrepresented in the publicly available data in PubChem. This “validation” dataset is likely to contain some misclassified compounds, as in the original training set. We correctly predicted the agonist and antagonist activity of over 340 compounds (>85%) using RFC predictor trained on the filtered dataset and over 350 (>88%) using XGBoost trained on the full dataset, demonstrating the power of this new *in silico* method.

Discussion

An observation that on average agonists are larger and more hydrophobic could potentially be used to distinguish them from antagonists. On the other hand, endogenous agonists adrenaline and noradrenaline are small and water soluble, suggesting that size and lipophilicity are not an intrinsic prerequisite of all agonists. An alternative explanation is that drug discovery efforts have focused on developing orally available (and therefore hydrophilic) β_1AR antagonists for the treatment of heart diseases. In contrast, the β_2AR agonists formulated for the treatment of asthma are delivered to the lungs via inhalation with higher hydrophobicity increasing their duration of action at the target tissue. This fits well with the observed correlation between the hydrophobicity and size and reflects a historical bias in the publicly available data. It is therefore clear that the pharmacological action of ligands on GPCRs is far more specific than a simple function of their size or hydrophobicity.

Studying the predicted ligand binding poses of thousands of ligands in the β_2AR binding pocket allowed us to identify the key ligand-receptor interactions which dictate a molecule’s propensity to cause agonism or antagonism. The structurally diverse nature of the test set allowed us to identify several “hot spots" mediating the agonism or antagonism of ligands acting on β_2AR. Agonism was mediated by residues in TM2 and TM5, and further facilitated by residues in TM6 and TM7. It is entirely plausible that certain ligands can successfully pull these TM regions together causing receptor

activation in the process. In contrast, our data suggest that antagonism is mediated by the interaction of ligands with W286$^{6.48x48}$, the so called toggle switch, that has long been proposed to play a key role in the activation of GPCRs[25,26]. The second mediator of antagonism is Y316$^{7.43x42}$ involved in the so called 3-7 lock that has previously been identified as important for GPCR activation[27]. Engaging these key residues in the ligand binding pocket likely prevent the conformational rearrangements necessary for activation of the receptor.

While the assembled data and our predictive models only classify compounds as agonists or antagonist, the pharmacological activity of compounds covers a spectrum from a very strong antagonist (aka inverse agonist) to that of a very strong agonist (aka full agonist). Another class of GPCR ligands, so called biased ligands, changes the balance between activating G protein and arrestin signalling pathways, with a potential to increase their therapeutic benefits[5,28]. It is likely that such partial and biased ligands would also show a district AIF of interactions they form with the receptor, and ML predictors could be developed to identify them in the in silico screen. However, a large experimental dataset of partial or biased agonists would be needed to explore this hypothesis, ideally collected in a uniform screen to minimise experimental and interpretational bias. The analysis of the learning performance of RFC and XGBoost classifiers (Figure S4) suggest that reasonable performance is achieved with this model with a limited dataset (ca 300-450 compounds), although further increases in the dataset size results in improved performance.

Our methodology can be readily applied to any receptor (or drug target) for which an extensive set of ligands has been developed and characterised, and where *in silico* docking experiments can be performed. This can include data already in the public domain or through examining the results of an *in house (eg, commercial)* drug-target screening campaigns. The advantage here is that in many cases the same signalling assay will have been used to profile all the compounds, improving the consistency of the dataset. This would allow the relative importance of each atom-atom interaction to be assessed as a modifier of signalling output. Also, it may be possible to isolate functional readouts (eg β–arrestin versus G protein) and therefore make predictions about functional bias. Further tantalising possibilities include the use of automated internet meta search of publications and patents to assemble such datasets if they are not available yet.

Being able to predict agonist or antagonist activity significantly increases the value of *in silico* docking campaigns. Importantly, it opens doors to a more rational engineering of ligands with improved and optimised pharmacological properties – facilitating the design of new ligands not present in the large virtual test libraries and thus opening up the chemical space many orders of magnitude above that of the largest virtual libraries available.

Conclusions

These results strongly support the hypothesis that the interatomic interactions between the receptor and its ligands are central to differentiate between their agonist and antagonist effects at the β_2AR. The overview obtained of the interatomic interactions between receptor and ligand which predicts an action enables the pharmacological classification of new compounds with previously unseen scaffolds. The growth of GPCR ligand databases provides a rich data source to facilitate the application of this approach to other GPCRs, while conceptually this approach could be applied to any drug target.

The ability to predict the pharmacological action of a ligand based on its ligand binding pose will significantly advance drug discovery projects contributing to a reduction of attrition during drug development. The tools presented have the potential to focus the efforts of chemists proposing new candidate molecules based on existing scaffolds and offers the opportunity to identify completely new scaffolds that may be more amendable to modifications from large scale docking experiments, thus opening up the chemical space many orders of magnitude above that of the largest virtual libraries available. They also have the potential to provide a rational and cost effective route to predict necessary modifications of existing drugs that are ineffective on receptors affected by genetic variations, paving the way towards the development of personalized medicines.

Acknowledgements

We thank Michael Stokes, Jill Baker, Uwe Grether, Arne Rufer and Wolfgang Guba for valuable advice and stimulating discussions. MJR was funded by COMPARE. BAM, CRH and SAC are funded by the MRC IMPACT doctoral programme. DAS is funded by the Roche Postdoctoral Fellowship. PH is funded by an international fellowship awarded by the Office of Educational Affairs, Thailand. AS is funded by the Study Abroad Postgraduate Education Scholarship (YLSY) awarded by the Republic of Turkey Ministry of National Education. LBR was funded by EMBO short term fellowship and the COST action 18133 ERNEST short term scientific mission fellowship. MJS was funded by the BBSRC Doctoral Training Programme, University of Nottingham and COMPARE.

References

1 Hauser, A. S. *et al.* Pharmacogenomics of GPCR Drug Targets. *Cell*, doi:10.1016/j.cell.2017.11.033 (2017).

2 Shore, S. A. & Moore, P. E. Regulation of beta-adrenergic responses in airway smooth muscle. *Respir Physiol Neurobiol* **137**, 179-195, doi:10.1016/s1569-9048(03)00146-0 (2003).

3 Cazzola, M., Calzetta, L. & Matera, M. G. β2-adrenoceptor agonists: current and future direction. *British Journal of Pharmacology* **163**, 4-17, doi:10.1111/j.1476-5381.2011.01216.x (2011).

4 Hilger, D., Masureel, M. & Kobilka, B. K. Structure and dynamics of GPCR signaling complexes. *Nature Structural & Molecular Biology* **25**, 4-12, doi:10.1038/s41594-017-0011-7 (2018).

5 Kenakin, T. & Miller, L. J. Seven transmembrane receptors as shapeshifting proteins: the impact of allosteric modulation and functional selectivity on new drug discovery. *Pharmacological reviews* **62**, 265-304, doi:10.1124/pr.108.000992 (2010).

6 Hilger, D. *et al.* Structural insights into differences in G protein activation by family A and family B GPCRs. *Science* **369**, doi:10.1126/science.aba3373 (2020).

7 Black, J. W., Crowther, A. F., Shanks, R. G., Smith, L. H. & Dornhorst, A. C. A New Adrenergic Betareceptor Antagonist. *Lancet* **1**, 1080-1081, doi:10.1016/s0140-6736(64)91275-9 (1964).

8 Frishman, W. H. beta-Adrenergic blockers: a 50-year historical perspective. *Am J Ther* **15**, 565-576, doi:10.1097/MJT.0b013e318188bdca (2008).

9 Cazzola, M., Rogliani, P. & Matera, M. G. Ultra-LABAs for the treatment of asthma. *Respiratory Medicine* **156**, 47-52, doi:10.1016/j.rmed.2019.08.005 (2019).

10 Cazzola, M., Rogliani, P., Stolz, D. & Matera, M. G. Pharmacological treatment and current controversies in COPD. *F1000Research* **8**, doi:10.12688/f1000research.19811.1 (2019).

11 de Graaf, C. & Rognan, D. Selective structure-based virtual screening for full and partial agonists of the beta2 adrenergic receptor. *J Med Chem* **51**, 4978-4985, doi:10.1021/jm800710x (2008).

12 Kooistra, A. J. *et al.* Function-specific virtual screening for GPCR ligands using a combined scoring method. *Scientific reports* **6**, 28288, doi:10.1038/srep28288 (2016).

13 Chan, H. C., Filipek, S. & Yuan, S. The Principles of Ligand Specificity on beta-2-adrenergic receptor. *Scientific reports* **6**, 34736, doi:10.1038/srep34736 (2016).

14 Warne, T., Edwards, P. C., Dore, A. S., Leslie, A. G. W. & Tate, C. G. Molecular basis for high-affinity agonist binding in GPCRs. *Science* **364**, 775-778, doi:10.1126/science.aau5595 (2019).

15 O'Boyle, N. M. *et al.* Open Babel: An open chemical toolbox. *J Cheminform* **3**, 33, doi:10.1186/1758-2946-3-33 (2011).

16 Trott, O. & Olson, A. J. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of computational chemistry* **31**, 455-461, doi:10.1002/jcc.21334 (2010).

17 Rasmussen, S. G. *et al.* Crystal structure of the beta2 adrenergic receptor-Gs protein complex. *Nature* **477**, 549-555, doi:10.1038/nature10361 (2011).

18 Ring, A. M. *et al.* Adrenaline-activated structure of beta2-adrenoceptor stabilized by an engineered nanobody. *Nature* **502**, 575-579, doi:10.1038/nature12572 (2013).

19 Staus, D. P. *et al.* Allosteric nanobodies reveal the dynamic range and diverse mechanisms of G-protein-coupled receptor activation. *Nature* **535**, 448-452, doi:10.1038/nature18636 (2016).

20 Jubb, H. C. *et al.* Arpeggio: A Web Server for Calculating and Visualising Interatomic Interactions in Protein Structures. *Journal of Molecular Biology* **429**, 365-371, doi:10.1016/j.jmb.2016.12.004 (2017).

21 Rodríguez-Espigares, I. *et al.* GPCRmd uncovers the dynamics of the 3D-GPCRome. *Nature Methods* **17**, 777-787, doi:10.1038/s41592-020-0884-y (2020).

22 Berman, H. M. The Protein Data Bank. *Nucleic Acids Research* **28**, 235-242, doi:10.1093/nar/28.1.235 (2000).

23 Tin Kam, H. in *Proceedings of 3rd International Conference on Document Analysis and Recognition* 278-282 (1995).

24 Chen, T. & Guestrin, C. in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785-794 (2016).

25 Shi, L. *et al.* β2Adrenergic Receptor Activation. *Journal of Biological Chemistry* **277**, 40989-40996, doi:10.1074/jbc.M206801200 (2002).

26 Visiers, I., Ballesteros, J. A. & Weinstein, H. in *G Protein Pathways Part A: Ribonucleases Methods in Enzymology* 329-371 (2002).

27 Trzaskowski, B. *et al.* Action of Molecular Switches in GPCRs - Theoretical and Experimental Studies. *Current Medicinal Chemistry* **19**, 1090-1109, doi:10.2174/092986712799320556 (2012).

28 Kenakin, T. & Christopoulos, A. Signalling bias in new drug discovery: detection, quantification and therapeutic impact. *Nature reviews. Drug discovery* **12**, 205-216, doi:10.1038/nrd3954 (2013).

Figures

Figure 1

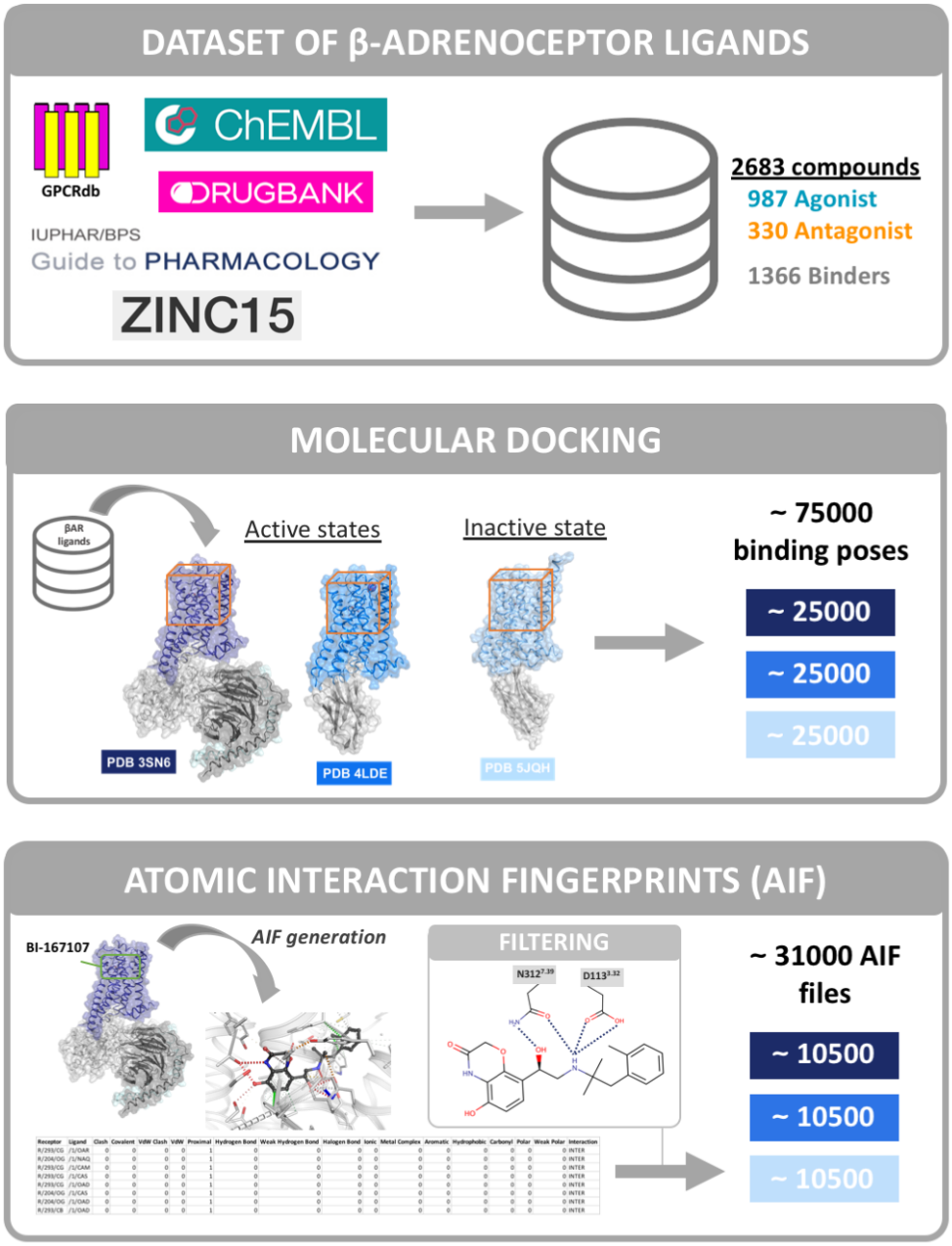
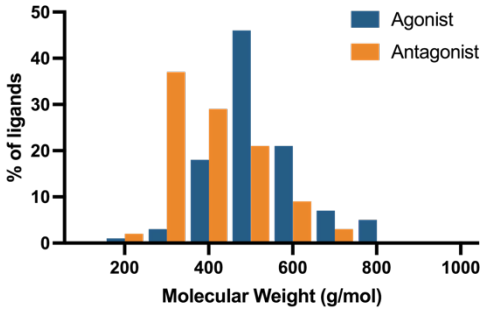


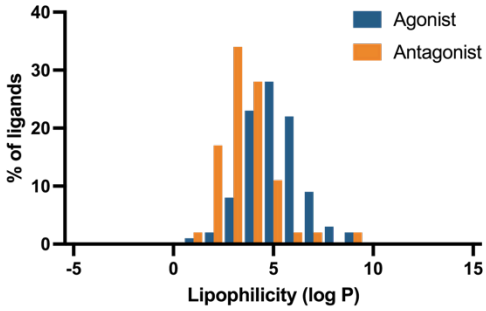
Figure 1. Workflow of the project. (A) Source of beta-adrenoceptor ligands available at open access repositories which comprise our 2683 compound dataset. (B) Molecular docking of test ligands to active and inactive β_2AR structures was performed using Autodock Vina. (C) Interatomic interaction fingerprint (AIF) calculations were made using Arpeggio. In case of the "filtered dataset", the generated AIFs were filtered based on the presence of ionic interactions with D113$^{3.32x32}$ and N312$^{7.39x38}$.

Figure 2

A



B



C

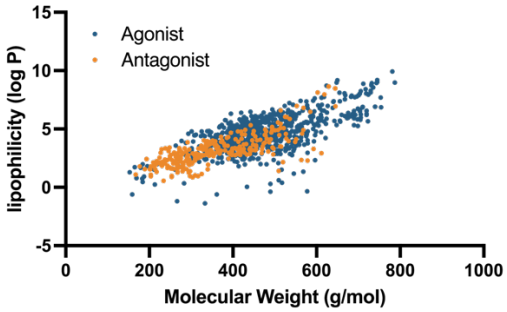
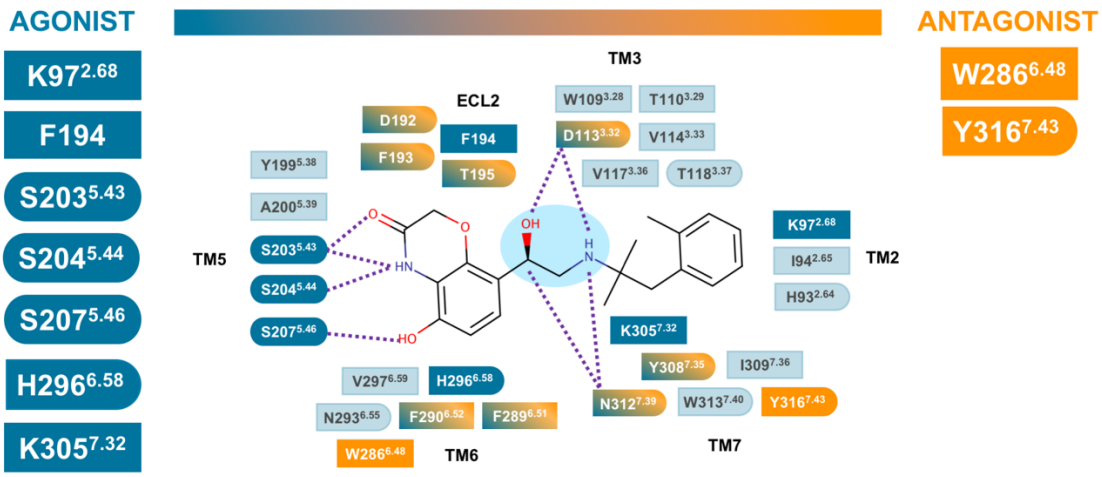


Figure 2. Physicochemical (PC) properties of the ligands (agonist in blue and antagonist in orange) predicted using OpenBabel software. (A) Molecular weight (MW) in g/mol, (B) lipophilicity (log P) and (C) correlation between lipophilicity and molecular weight. Spearman correlation coefficient is 0.62 for agonist and 0.76 for antagonists.

Figure 3

A) FILTERED DATASET



B) FULL DATASET

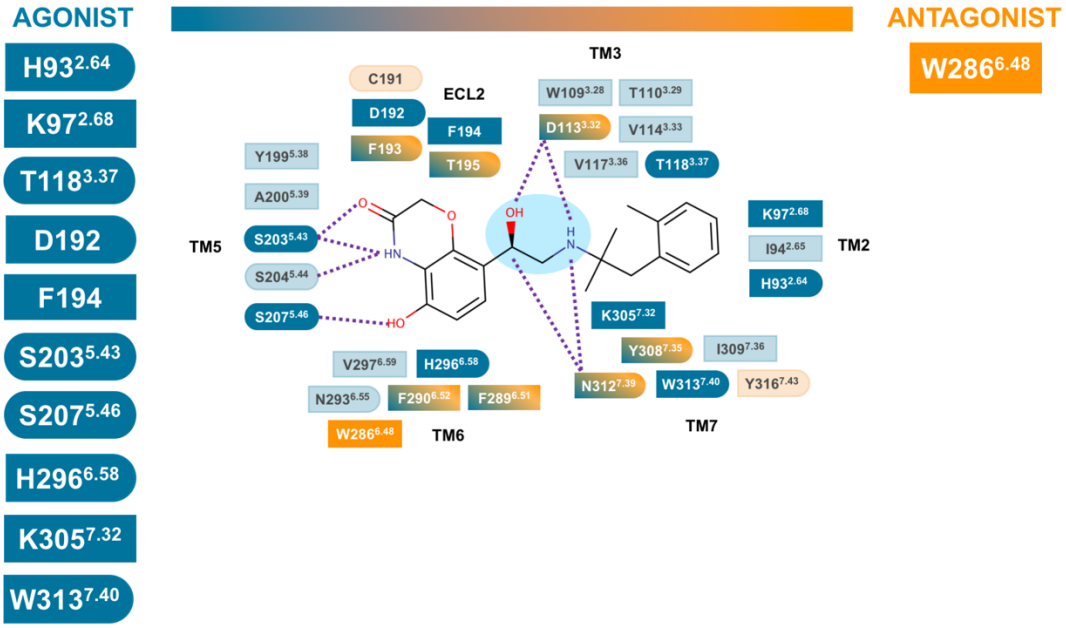
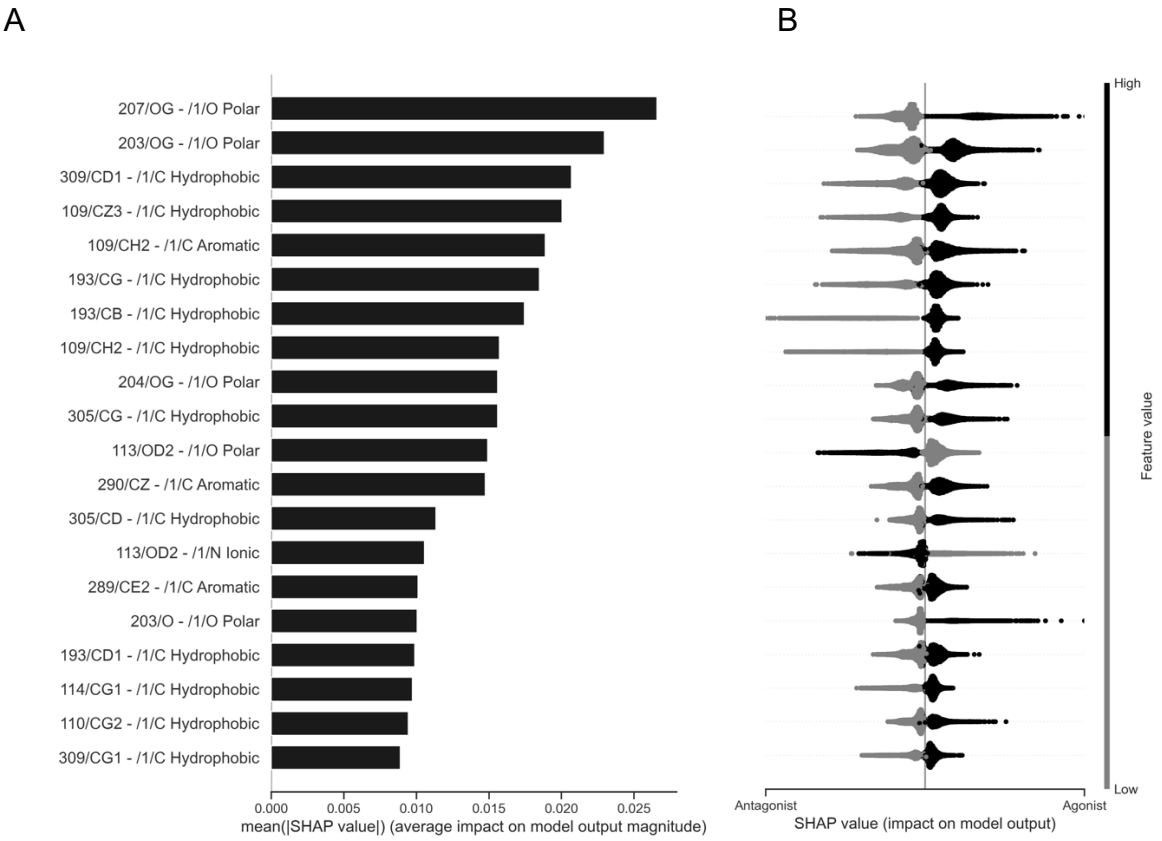


Figure 3. Schematic representation of the interactions predicted using the pairwise correlation approach. (A) filtered dataset and (B) full dataset. The type of interaction is summarised in squared shape for hydrophobic and aromatic contacts, round shape for the polar, ionic and hydrogen bond contacts, and the combination of both. The doted purple line represent ionic and/or hydrogen bond contacts. The ethanolamine moiety of the BI-167107 ligands is highlighted in light blue.

Figure 4


A
B
207/OG - /1/O Polar
203/OG - /1/O Polar
309/CD1 - /1/C Hydrophobic
109/CZ3 - /1/C Hydrophobic
109/CH2 - /1/C Aromatic
193/CG - /1/C Hydrophobic
193/CB - /1/C Hydrophobic
109/CH2 - /1/C Hydrophobic
204/OG - /1/O Polar
305/CG - /1/C Hydrophobic
113/OD2 - /1/O Polar
290/CZ - /1/C Aromatic
305/CD - /1/C Hydrophobic
113/OD2 - /1/N Ionic
289/CE2 - /1/C Aromatic
203/O - /1/O Polar
193/CD1 - /1/C Hydrophobic
114/CG1 - /1/C Hydrophobic
110/CG2 - /1/C Hydrophobic
309/CG1 - /1/C Hydrophobic
0.000
0.005
0.010
0.015
0.020
0.025
mean(|SHAP value|) (average impact on model output magnitude)
High
Feature value
Low
Antagonist
Agonist
SHAP value (impact on model output)

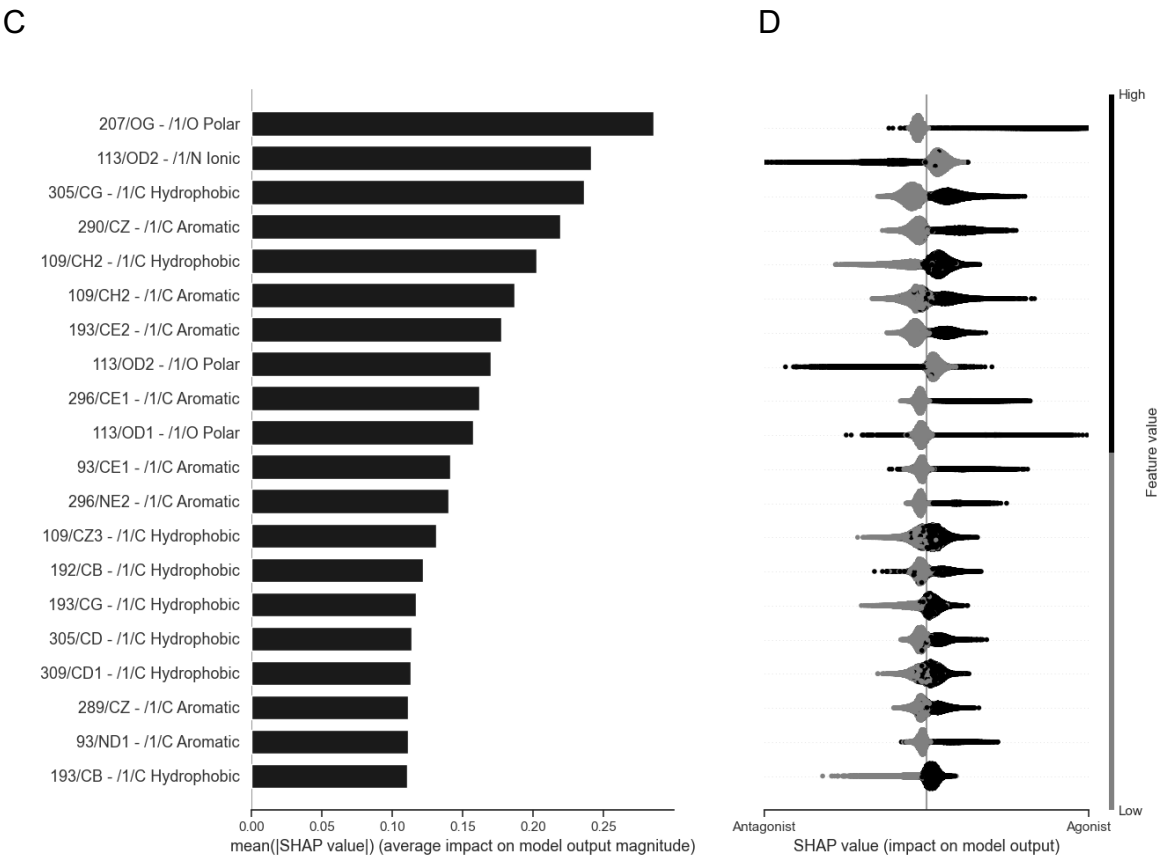
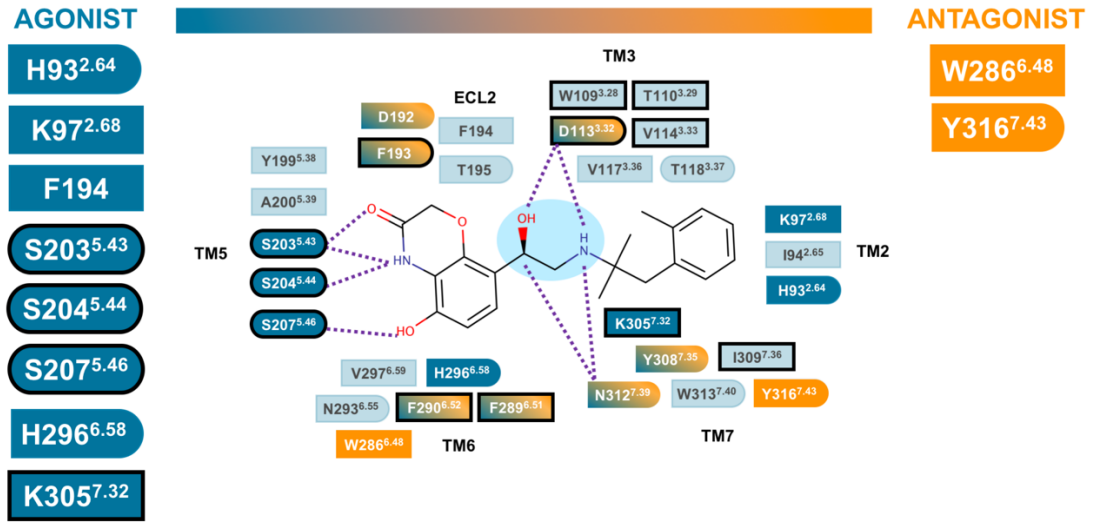


Figure 4. Feature importance of the RFC (A,B) and XGBoost ML (C,D) models applying SHAP value method. (A,C) The x-axis is the average magnitude change in model output when a feature is “hidden” from the model. Higher SHAP values indicate higher importance of the feature. (B,D) Local SHAP values per sample (each ligand pose) sorted by the mean absolute SHAP value method. Grey represents a value of 0, thus indicating the absence of a particular atomic interaction for a specific sample. Black represents a value of 1, thus indicating the presence of a particular atomic interaction for a specific sample. The x-axis shows how the presence or absence of an atomic feature increases or decreases the likelihood of a sample being classified as an antagonist. The data are plotted for all samples in the dataset, showing the distribution of importance values. The units of the x-axis using RFC and XGBoost are log odds.

Figure 5

A) FILTERED DATASET (RFC)



B) FULL DATASET (XGBoost)

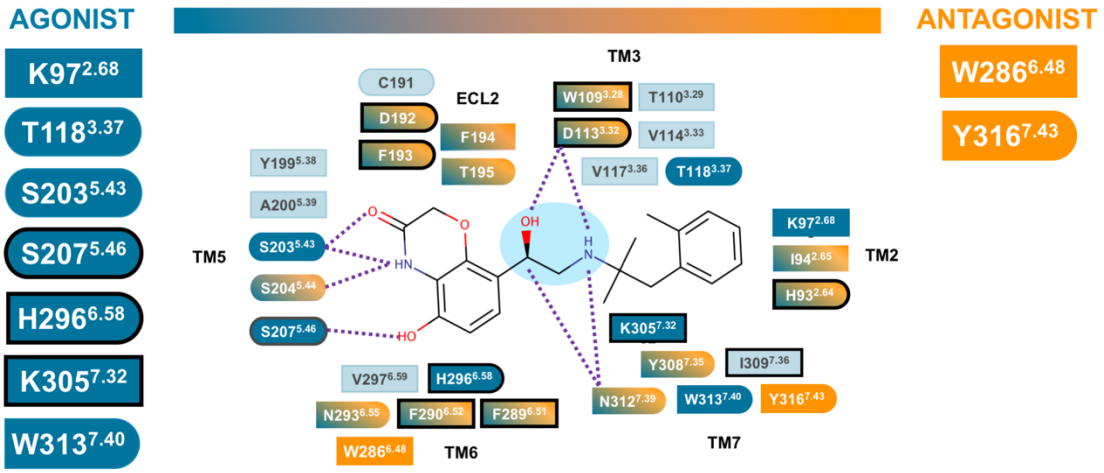
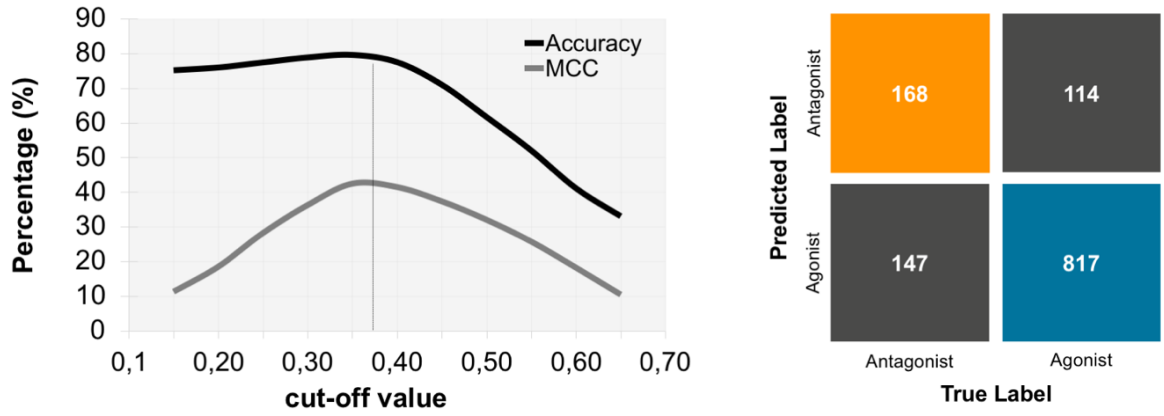


Figure 5. Schematic representation of the interactions for the machine learning approach. (A) RFC for the filtered dataset and (B) XGBoost for full dataset. The type of interaction is summarised in squared shape for hydrophobic and aromatic contacts, round shape for the polar, ionic and hydrogen bond contacts, and the combination of both. The dotted purple line represent ionic and/or hydrogen bond contacts. The black outlines represent the atomic interactions with higher feature importance. The ethanolamine moiety of the BI-167107 ligands is highlighted in light blue.

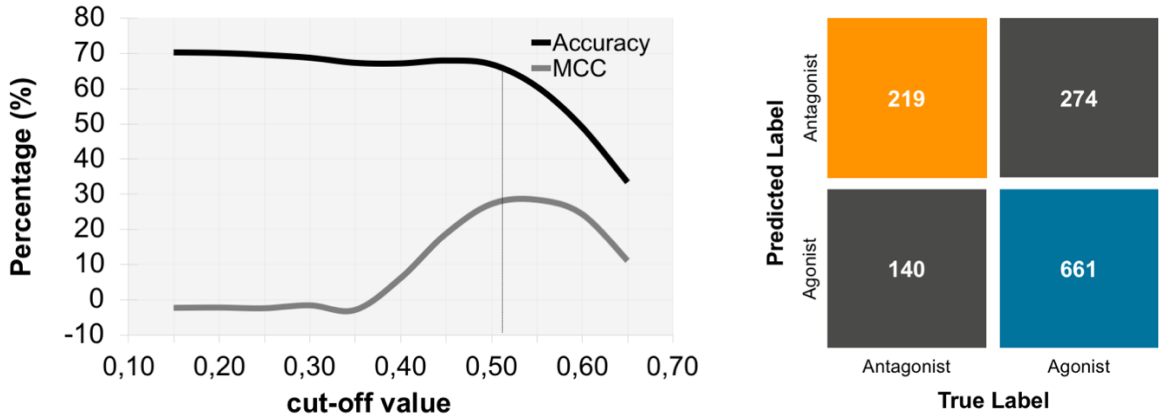
Supplementary data

Supplementary Figure 1

B. FILTERED DATASET



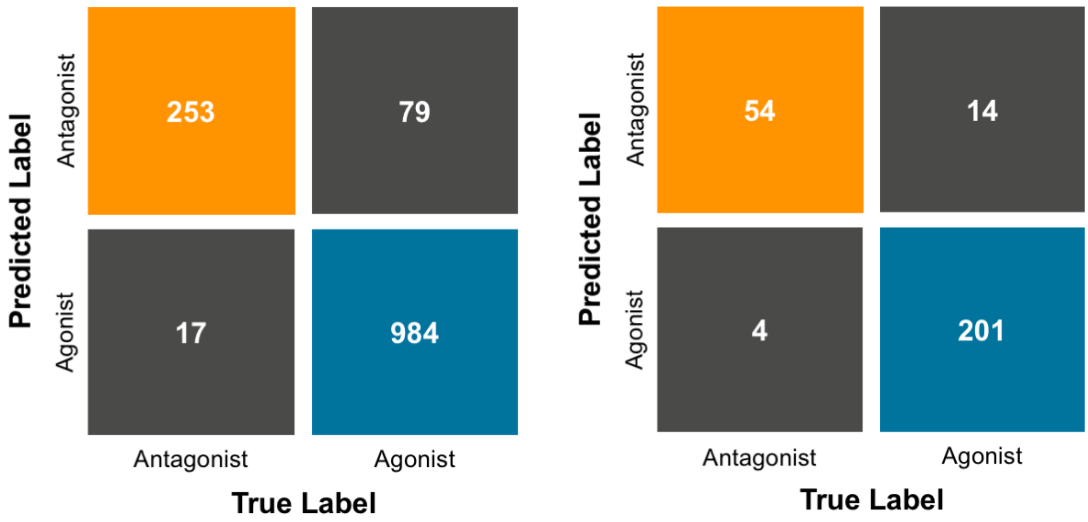
B. FULL DATASET



Supplementary Figure 1. Cut-off graph to obtain the maximum MCC value for the Pearson's correlation approach and the corresponding confusion matrix to differentiate agonist and antagonist. (A) Filtered dataset, right: cut-off and left: confusion matrix MCC = 0.37. (B) Full dataset, right: cut-off and left: confusion matrix MCC = 0.51.

Supplementary Figure 2

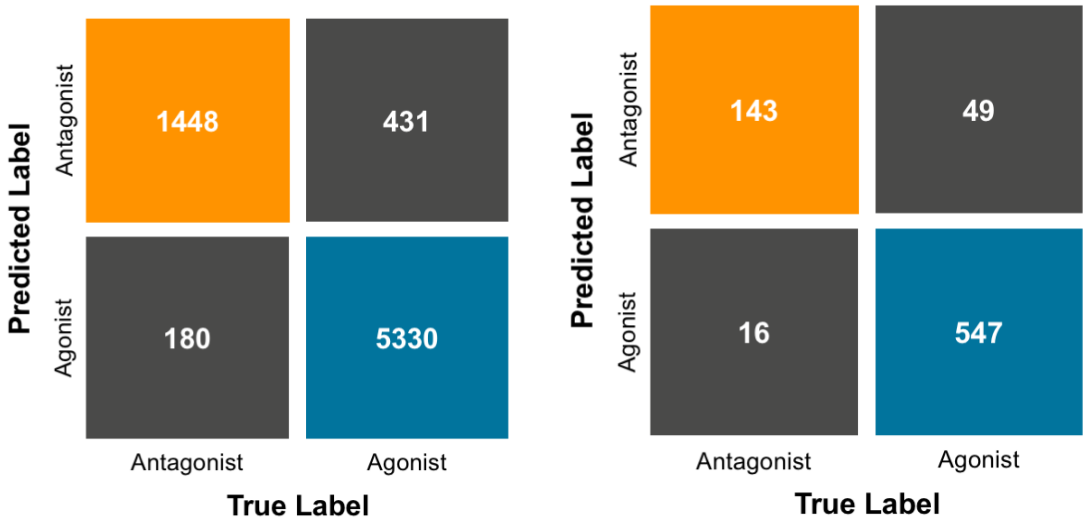
FILTERED DATASET



Supplementary Figure 2. Random Forest Classification Confusion Matrix for Cross-Validation (left) and Hold-out sets (right) for the filtered dataset. The left figure shows the median confusion matrix across all K-folds and repeats, the right shows the hold-out set. Y-axis shows the frequency of classes predicted by the model, across all samples. The x-axis shows the true known class label. Overall the confusion matrix provides insight into the ratios of false positive & negatives made by the model across samples for the filtered dataset.

Supplementary Figure 3

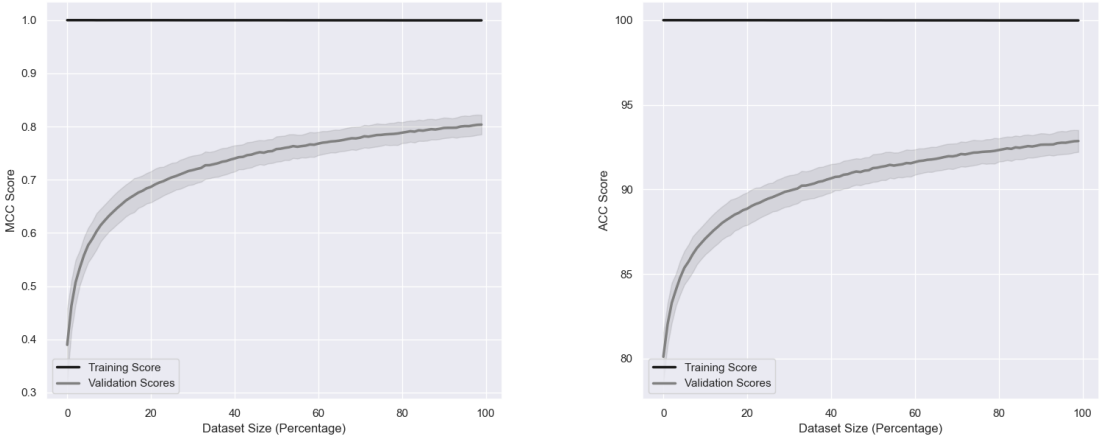
FULL DATASET



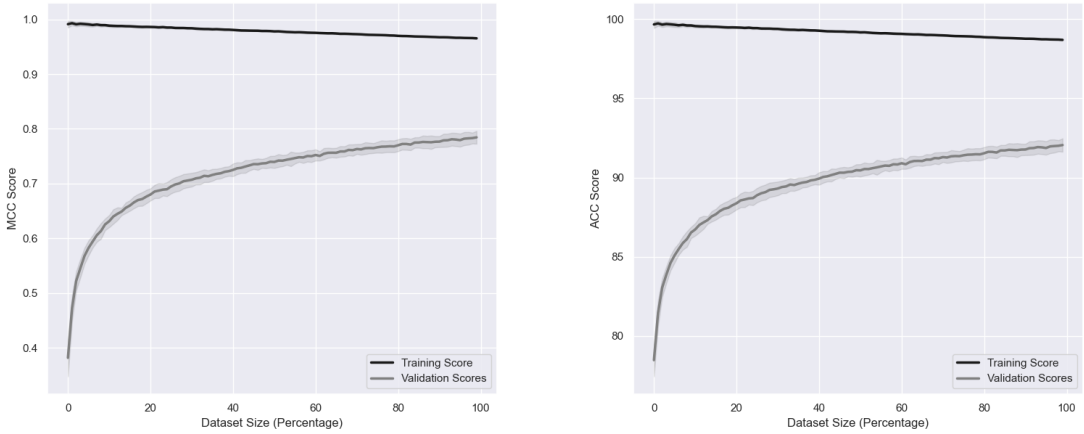
Supplementary Figure 3. XGBoost Confusion Matrix for Cross-Validation (left) and Hold-out sets (right) for the full dataset.

Supplementary Figure 4

A) FILTERED DATASET



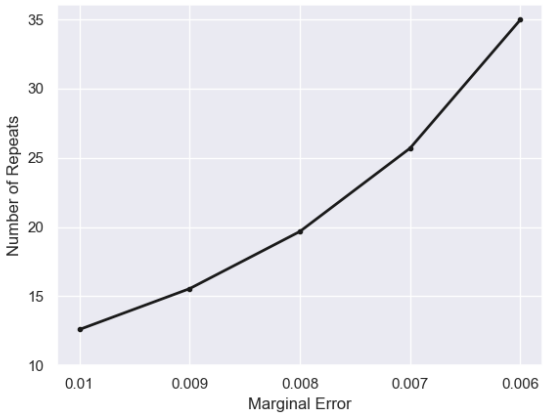
B) FULL DATASET



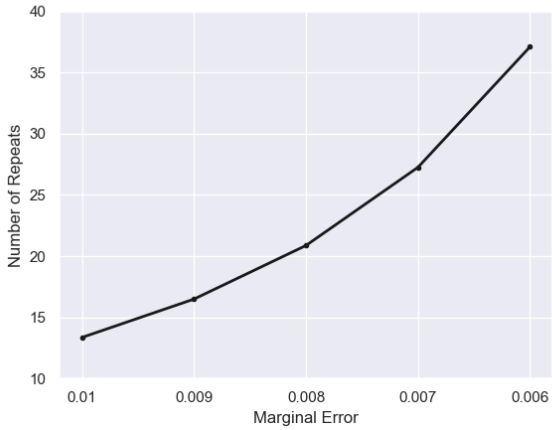
Supplementary Figure 4. Learning curves for the MCC Scores (left) and accuracy – ACC Scores (rights) for the (A) RFC model on the filtered dataset and the (B) XGBoost model on the unfiltered dataset.

Supplementary Figure 5

A



B



Supplementary Figure 5. Determining the optimal number of repeats required to estimate, with 99% confidence, true model performance to a specified level of precision. (A) RFC model on the filtered dataset and (B) XGBoost model on the unfiltered dataset. In this case, 13 repeats were sufficient for both models.

Additional files:

Supplementary Tables 1-5 are in included as an Excel file "SupplementaryTables_1-5.xlsx".

Methods

Dataset preparation

A dataset was compiled using the primary open access repositories GPCRdb[1,2], ChEMBL[3], ZINC[4], DrugBank[5] and Guide to Pharmacology[6]. This dataset yielded a total of 2643 unique β_2AR ligands, of which 1317 have reported pharmacological action, while 1326 compounds are binders with undetermined activity profiles. We classify ligands with known activity as either agonists (including partial and full agonists) or antagonists (including inverse agonists).

Each ligand was assigned an internal ID (ranging from 1 to 2643), and its corresponding SMILES string (line notation encoding its molecular structure) and pharmacological action (agonist/antagonist/binder) were retrieved from the relevant databases. The International Chemical Identifier key (InChIKey) was used as a unique identifier to distinguish between ligands across the dataset[7]. Both InChIKey and physicochemical properties appended for all compounds were acquired using the software Open Babel v3.1.1[8].

Protein structures and ligand preparation and Docking

The active-state protein coordinates were extracted from two crystal structures of human β_2AR bound to an ultrahigh-affinity agonist (BI167107) coupled with the Gs protein[9] or/and with a G protein-mimicking nanobody (Nb6B9)[10] from the Protein Data Bank (PDB code: 3SN6 and 4LDE, respectively). The inactive-state protein coordinates were extracted from the human β_2AR bound to the inverse agonist carazolol (PDB code: 5JQH)[11].

Receptor structures were aligned to use the same grid box of $22 \times 22 \times 32$ Å at the orthosteric binding site, protonated and charged, yielding a protein input file for subsequent docking experiments using UCSF Chimera[12].

The SMILES representation of ligands along with their internal ID were protonated and converted to a spatial data file (SDF) and pdbqt formats using Obabel.

The semi-flexible molecular docking was carried out using the software AutoDock Vina[13] and generated up to 10 poses for every compound. In total, 2643 compounds were docked in three β-adrenoceptor structures, yielding almost 27000 docking poses.

Interaction fingerprint calculations and filtering

The inter-atomic receptor-ligand interaction fingerprints (AIFs) were calculated for all docking poses generated for each compound using the software Arpeggio[14] executed in Docker environment[15], a software container platform. This method accounts for the presence of up to 15 subtypes of interatomic interactions, classified by atom type, distance and angle constraints. The output was presented as binary values, with a 1 denoting the presence of a particular defined interaction and 0 indicating an absence.

A Python script was written to filter the Arpeggio results (to generate the "filtered dataset") by imposing minimum constraints that enforced certain features deemed essential for β_2AR ligand binding, which it eliminated all irrelevant binding poses (around 50% rows). Criteria important for binding were based on prior knowledge derived from the literature, in particular the presence of the ionic/polar interaction between D113$^{3.32}$ and N312$^{7.39}$ with the ethanolamine moiety of the ligands.

Generation of interaction matrix

A Python script was written to process each docking pose to generate a single *MxN* matrix for each PDB, where *M* are ligand poses (samples as 'ligand internal ID_docking pose number') and *N* are the specific atomic interaction (features as 'receptor residue number/interacting atom – ligand interacting atom and interaction type' (e.g. 'lig 752_04' and '301/O - N Polar' respectively)), present in the whole ligand set. The value 1 corresponds to the occurrence of a particular type of interaction and 0 to the non-occurrence of a particular type of interaction. As the AIF files generated by Arpeggio only contained the interaction present for a particular docking pose, the imputation of missing data was handled by setting any undefined (NaN) values to 0. Finally, we excluded the 3 subtypes of interactions reported by Arpeggio: Clash, VdW Clash and Proximal from subsequent analysis as they provided very little information but represented around 60% of the columns. We also included pharmacological action

label (“agonist” or “antagonist”, if known, or “binder” if not known) for each ligand binding pose in the same data table as an additional column.

Descriptive statistical analysis

A descriptive statistical analysis of the frequency of the interatomic interactions was performed using a Python script. In this manner, the most frequently occurring features (those observed in at least 10% of all docking poses) contributing to agonism and antagonism were clearly identified across all interaction types and collated for further analysis. This resulted in the reduction of the number of features from ca 1,100 to ca 100.

Subsequently, we computed pairwise correlation between the columns representing atomic interactions and the column representing pharmacological action using Pearson’s correlation coefficient (r) method. The resulting value r for each interaction (feature) reflects how well it is correlated with the pharmacological action (agonism or antagonism). Plots, graphs and tables were generated with Excel, and statistical significance was determined using an unpaired t-test using Prism 8.

Machine learning dataset preparation

The dataset was then randomly shuffled and split, via stratification, into cross-validation and final hold-out datasets. The cross-validation set was used for training and validation during hyperparameter optimisation. The hold-out dataset comprised 2% of the original dataset and allowed us to gauge whether the validation scores were good estimations of model performance when generalising to unseen data. The hold-out set was not used during any training or optimisation procedures.

Model Selection

Performance Metric

For the filtered dataset the Random Forest classifier was used. For the unfiltered dataset both Random Forest and XGBoost classifiers were used. Matthews correlation coefficient (MCC) was used as the performance metric for all models[16]. The MCC metric is defined as follows:

$$MCC = \frac{TP * TN - FP * FN}{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}$$

Where TP is the number of true positives, TN the number of true negatives, FP the number of false positives and FN the number of false negatives. The MCC for binary classification weights both positive and negative classes equally, whilst also being robust to severe class imbalances. A value of +1 indicates a perfect positive correlation, that is a total agreement between prediction and observation. An MCC score of 0 indicates no correlation, that is the classifier performs no better than a random coin flip. Finally, -1 indicates a perfect negative correlation, that is a total disagreement between prediction and observation.

Model Performance Estimation

Model performance was validated using repeated-stratified-k-fold cross-validation. Cross validation entails splitting the dataset into k equally sized partitions, termed folds. One of the folds is extracted and used for validating a model on unseen data. The remaining folds are then used to train the model. This process is then repeated using each of the k folds as the validation set. The optimal model is the one which has the best performance on average across all k folds. Cross validation generally provides a less optimistic estimation of model generalizability on unseen data, which is finally tested on the holdout set. Due to class imbalances in the data, stratification is used to ensure that the original distribution of classes is maintained in each fold, thus preventing any fold from being populated by a single class[17]. Model estimation can be noisy and so by performing cross-validation over many repeats one obtains a more precise estimation of true model performance. Bootstrap resampling was used to estimate model uncertainty[18] . Confidence intervals were calculated with respect to a 99% confidence level. Bayesian hyperparameter optimisation (BHO) was utilised to determine high performing model parameter configurations when tested on unseen data. BHO was set to maximize the mean MCC across K-folds and repeats, model uncertainty was then calculated using optimised models only.

Random Forest Classifier Hyper Parameters

Hyperparameter Name	Hyperparameter Value
Criterion	Gini
Estimators	2000
Max Depth	5
Max-Features	5
Splitter	Best
Minimum Samples Split	2
Minimum Samples Leaf	1
Minimum weighted fraction Leaf	0
Maximum Leaf Nodes	Unlimited
Minimum Impurity Split	None
Minimum Impurity Decrease	None

XGBoost Hyper Parameters

Hyperparameter Name	Hyperparameter Value
Eta	0.4306
gamma	0.2458
Learning Rate	0.05873
Max Delta Step	7
Max Depth	8
Minimum Child Weight	1.246
Number of Estimators	1150
Scale Positive Weight	1
Subsample	0.7532

All other hyperparameters for XGBoost that are not specified were kept at their default values according to the XGBoost API guide (https://xgboost.readthedocs.io/en/latest/parameter.html)

Model Feature Importance analysis

The most important atomic interactions, for classifying agonist or antagonist ligands, were identified using the Shapley Additive Explanations (SHAP) method[19]. Shapley values are based upon coalition game theory and inform one how to fairly distribute

the prediction of a model among the features. The Shapley value for one feature is the average marginal contribution of a feature value across all the possible combinations of features. More concretely the Shapley value assigns an importance to each feature by calculating the effect on model prediction when including a particular feature compared to the model prediction when the feature is withheld. Mathematically this can be formalised as

$$\phi_i = \sum_{S \subseteq M \setminus i} \frac{|S|!(|M|-|S|-1)!}{|M|!} [f(S \cup i) - f(S)],$$

where S refers to a subset of features that does not contain the feature for which we are calculating ϕ_i. $S \cup i$ is the subset that contains features in S and feature i. Finally, $S \subseteq M/i$ represents all sets S that are subsets of the total set of features M, excluding feature i. The computation time increases exponentially with the number of features; thus we used the TreeSHAP algorithm that approximates SHAP values for tree-based machine learning models in polynomial time[20]. The main motivations for using the SHAP feature importance method over other popular methods, such as Gini and Permutation methods, is due to the following:

Consistency: The Gini feature importance method is susceptible to producing inconsistent feature importances that are biased to the specific ordering of features specified by their position, as split nodes, in the tree. TreeSHAP method is equivalent to averaging differences in model predictions over all possible orderings of the features and thus does not suffer from such inconsistencies.

Granular Interpretability: Although permutation importance is not biased to the specific structure of decision trees it only provides a global understanding of the most important features. With TreeSHAP, observations get their own set of SHAP values and therefore we can understand feature importance on a per sample basis.

Determining The Optimal Number of Repeats

There is an exponential relationship between the number of times one has to repeat Bootstrap or Cross-Validation and the level of precision to within which one would like to measure true model performance. This leads to a trade-off between the precision and time complexity of model performance estimation. We thus estimate the optimal

number of repeats to use for Bootstrap and cross-validation resampling methods to an acceptable level of precision as:

$$n = z_a^2 \left(\frac{\sigma^2}{\delta^2} \right)$$

where z is the ordinate on the Normal distribution curve that corresponds to a particular level of confidence we have in our estimation, denoted α. σ is the population standard deviation and δ is the specified precision of the estimate. We estimate the population standard deviation via repeated bootstrap resampling, thus each estimate of the number of repeats is specific to the variance of each model and its hyperparameter configuration (https://www.itl.nist.gov/div898/handbook/ppc/section3/ppc333.htm).

A precision of 1% (Marginal Error = 0.01) was selected for all resampling methods (Supplementary Figure 5). Therefore, a minimum of 13 repeats for both the RFC and XGBoost were used during cross-validation and bootstrap resampling.

References

1 Pandy-Szekeres, G. *et al.* GPCRdb in 2018: adding GPCR structure models and ligands. *Nucleic Acids Res*, doi:10.1093/nar/gkx1109 (2017).

2 Hauser, A. S., Attwood, M. M., Rask-Andersen, M., Schioth, H. B. & Gloriam, D. E. Trends in GPCR drug discovery: new agents, targets and indications. *Nat Rev Drug Discov* **16**, 829-842, doi:10.1038/nrd.2017.178 (2017).

3 Mendez, D. *et al.* ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Research* **47**, D930-D940, doi:10.1093/nar/gky1075 (2019).

4 Irwin, J. J. & Shoichet, B. K. ZINC – A Free Database of Commercially Available Compounds for Virtual Screening. *Journal of Chemical Information and Modeling* **45**, 177-182, doi:10.1021/ci049714+ (2005).

5 Wishart, D. S. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Research* **34**, D668-D672, doi:10.1093/nar/gkj067 (2006).

6 Armstrong, J. F. *et al.* The IUPHAR/BPS Guide to PHARMACOLOGY in 2020: extending immunopharmacology content and introducing the IUPHAR/MMV Guide to MALARIA PHARMACOLOGY. *Nucleic Acids Research*, doi:10.1093/nar/gkz951 (2019).

7 Heller, S., McNaught, A., Stein, S., Tchekhovskoi, D. & Pletnev, I. InChI - the worldwide chemical structure identifier standard. *Journal of Cheminformatics* **5**, doi:10.1186/1758-2946-5-7 (2013).

8 O'Boyle, N. M. *et al.* Open Babel: An open chemical toolbox. *J Cheminform* **3**, 33, doi:10.1186/1758-2946-3-33 (2011).

9 Rasmussen, S. G. *et al.* Crystal structure of the beta2 adrenergic receptor-Gs protein complex. *Nature* **477**, 549-555, doi:10.1038/nature10361 (2011).

10 Ring, A. M. *et al.* Adrenaline-activated structure of beta2-adrenoceptor stabilized by an engineered nanobody. *Nature* **502**, 575-579, doi:10.1038/nature12572 (2013).

11 Staus, D. P. *et al.* Allosteric nanobodies reveal the dynamic range and diverse mechanisms of G-protein-coupled receptor activation. *Nature* **535**, 448-452, doi:10.1038/nature18636 (2016).

12 Pettersen, E. F. *et al.* UCSF Chimera--a visualization system for exploratory research and analysis. *Journal of computational chemistry* **25**, 1605-1612, doi:10.1002/jcc.20084 (2004).

13 Trott, O. & Olson, A. J. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of computational chemistry* **31**, 455-461, doi:10.1002/jcc.21334 (2010).

14 Jubb, H. C. *et al.* Arpeggio: A Web Server for Calculating and Visualising Interatomic Interactions in Protein Structures. *Journal of Molecular Biology* **429**, 365-371, doi:10.1016/j.jmb.2016.12.004 (2017).

15 Merkel, D. Docker: lightweight Linux containers for consistent development and deployment. *Linux J.* **2014**, Article 2 (2014).

16 Matthews, B. W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure* **405**, 442-451, doi:10.1016/0005-2795(75)90109-9 (1975).

17 Vanwinckelen, G. & Blockeel, H. in *Belgian-Dutch Conference on Machine Learning (BeneLearn).* (eds Bernard Manderick, Michaël Rademaker, & Willem Waegeman) 39-44 (BeneLearn 2012: Proceedings of the 21st Belgian-Dutch Conference on Machine Learning).

18 Efron, B. & Tibshirani, R. *An introduction to the bootstrap*. (Chapman & Hall, 1993).

19 Lundberg, S. & Lee, S.-I. *A Unified Approach to Interpreting Model Predictions*. (2017).

20 Lundberg, S. M. *et al.* From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence* **2**, 56-67, doi:10.1038/s42256-019-0138-9 (2020).