UNIVERSIDADE FEDERAL DE MINAS GERAIS
Programa Interunidades de Pós-Graduação em Bioinformática

João Paulo Linhares Velloso

In silico discovery of GPCR ligands using graph-based signatures and auxiliary features

Belo Horizonte

2022

João Paulo Linhares Velloso

In silico discovery of GPCR ligands using graph-based signatures and auxiliary features

Versão Final

Tese apresentada ao Programa Interunidades de Pós-Graduação em Bioinformática da Universidade Federal de Minas Gerais, como requisito parcial para a obtenção do grau de Doutor em Bioinformática.

Orientador: Prof. Dr. Douglas Eduardo Valente Pires
Coorientador: Prof. Dr. David Benjamin Ascher

Belo Horizonte

2022

043

Velloso, Joao Paulo Linhares.
In silico discovery of GPCR ligands using graph-based signatures and auxiliary features [manuscrito] / Joao Paulo Linhares Velloso. – 2022.
129 f. : il. ; 29,5 cm.

Orientadores: Prof. Dr. Douglas Eduardo Valente Pires e Prof. Dr. David Benjamin Ascher.
Tese (doutorado) – Universidade Federal de Minas Gerais, Instituto de Ciências Biológicas. Programa de Pós-Graduação em Bioinformática.

1. Bioinformática. 2. Aprendizado de Máquina. 3. Desenvolvimento de Medicamentos. 4. Receptores Acoplados a Proteínas-G. I. Pires, Douglas Eduardo Valente. II. Ascher, David Benjamin. III. Universidade Federal de Minas Gerais. Instituto de Ciências Biológicas. IV. Título.

CDU: 573:004

Ficha catalográfica elaborada pela bibliotecária Fabiane C M Reis – CRB 6 - 2680

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS BIOLÓGICAS
PROGRAMA DE PÓS-GRADUAÇÃO EM BIOINFORMÁTICA

ATA DA DEFESA DE TESE

JOÃO PAULO LINHARES VELLOSO

Às dezessete horas do dia **27 de abril de 2022**, reuniu-se, através de videoconferência, a Comissão Examinadora de Tese, indicada pelo Colegiado do Programa, para julgar, em exame final, o trabalho intitulado: **"In silico discovery of GPCR ligands using graph-based signatures and auxiliary features"**, requisito para obtenção do grau de Doutor em **Bioinformática.** Abrindo a sessão, o Presidente da Comissão, **Dr. Douglas Eduardo Valente Pires**, após dar a conhecer aos presentes o teor das Normas Regulamentares do Trabalho Final, passou a palavra ao candidato, para apresentação de seu trabalho. Seguiu-se a arguição pelos Examinadores, com a respectiva defesa do candidato. Logo após, a Comissão se reuniu, sem a presença do candidato e do público, para julgamento e expedição de resultado final. Foram atribuídas as seguintes indicações:

Professor(a)/Pesquisador(a)	**Instituição**	**Indicação**
Dr. Douglas Eduardo Valente Pires	Fundação Oswaldo Cruz	**Aprovado**
Dr. Lucas Bleicher	Universidade Federal de Minas Gerais	**Aprovado**
Dra. Rafaela Salgado Ferreira	Universidade Federal de Minas Gerais	**Aprovado**
Dr. Rubens Lima do Monte Neto	Fundação Oswaldo Cruz	**Aprovado**
Dr. Wandré Nunes de Pinho Veloso	Universidade Federal de Itajubá	**Aprovado**

Pelas indicações, o candidato foi considerado: **Aprovado**

O resultado final foi comunicado publicamente ao candidato pelo Presidente da Comissão. Nada mais havendo a tratar, o Presidente encerrou a reunião e lavrou a presente ATA, que será assinada por todos os membros participantes da Comissão Examinadora.

Belo Horizonte, 27 de abril de 2022.

Documento assinado eletronicamente por **Rubens Lima do Monte Neto**, **Usuário Externo**, em 27/04/2022, às 17:52, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Wandré Nunes de Pinho Veloso**, **Usuário Externo**, em 27/04/2022, às 19:56, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Douglas Eduardo Valente Pires**, **Usuário Externo**, em 27/04/2022, às 21:01, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Rafaela Salgado Ferreira**, **Professora do Magistério Superior**, em 28/04/2022, às 14:20, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Lucas Bleicher**, **Professor do Magistério Superior**, em 28/04/2022, às 14:35, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufmg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **1386436** e o código CRC **049629B7**.

Referência: Processo nº 23072.221589/2022-53 SEI nº 1386436

Acknowledgements

I would like to thank the Universidade Federal de Minas Gerais for supporting this PhD research, and also for providing me the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior scholarship. I also would like to thank my supervisors, Professor Douglas Pires and Professor David Ascher for carefully, and with patience, guiding me throughout all steps I gave in this journey. Not only that, but I am very grateful for all the teachings, discussions, and assistance they provided these last years. I am also grateful for the support I had from the Instituto René Rachou, providing me with all the infrastructure I needed for developing my computational experiments.

Throughout this PhD, I was also backed up by wonderful friends and colleagues. They were essential for my success and made my life easier, through academic help or providing emotional support. I would like to mention and thank here Amanda, Rubens, Fausto, Gabriel, Fabiano, Anna Carol, Joicy, Alex, Pâmela and Juliana Assis. I also would like to thank all professors that accepted being part of my PhD evaluation committee.

Finally, I would like to thank my parents, Roberto, Edilene, my grandmothers Any, and Zélia, and my brother Marco for always being present and available whenever and wherever I needed. They are the best gifts I could have ever received.

Resumo

Os receptores acoplados a proteína G (GPCR) são cruciais para muitos processos fisiológicos vitais, incluindo controle da divisão e proliferação celular, regulação do transporte de íons, modulação sinapse nervosa, homeostase, modulação e modificação da morfologia celular. Eles também estão envolvidos em muitos processos patológicos, como Alzheimer e Parkinson, distúrbios cardiovasculares, asma, depressão e diabete. Dada a sua importância biológica, mais de um terço dos medicamentos aprovados pela FDA têm como alvo esses receptores. No entanto, o desenvolvimento de fármacos para GPCRs passa por altas taxas de fracasso, com baixa eficácia *in vivo* sendo o principal contribuinte nesse processo. Isso resulta em apenas 7% de todos os medicamentos (incluindo outros receptores) em estudos de fase I sendo comercializados. Esta tese se concentrou no desenvolvimento de modelos de aprendizado de máquina capazes de prever a bioatividade de pequenas moléculas ao interagir com GPCRs. Pretendemos com essas ferramentas apoiar a descoberta de novos fármacos. Os modelos desenvolvidos (compõe o servidor web pdCSM-GPCR) baseiam-se em derivar uma série de assinaturas moleculares de ligantes conhecidos, associando essas assinaturas a bioatividade e modelando essas questões como problemas de regressão, sem a necessidade de informação estrutural do receptor. Devido a esta característica, a mesma abordagem pode ser usada para quaisquer GPCRs que já tenham sido avaliadas através triagem para ligantes, e também para outros alvos importantes, incluindo quinases e canais iônicos controlados por ligantes. Nossos modelos compõem o recurso computacional mais abrangente para previsão da bioatividade de GPCR até o momento, e inclui também suporte para o desenvolvimento de medicamentos para GPCRs órfãos. Nossa abordagem alcançou correlações de Pearson de até 0,89, por meio de validação cruzada de 10 vezes e em testes cegos. Superamos significativamente os métodos anteriores. O pdCSM-GPCR foi disponibilizado gratuitamente por meio um servidor web http://biosig.unimelb.edu.au/pdcsm_gpcr. Também investigamos as propriedades de pequenas moléculas com alta afinidade por GPCRs a fim de identificar determinantes moleculares de reconhecimento. Em geral, ligantes potentes possuem fragmentos contendo nitrogênio e anéis aromáticos, características comuns em ligantes em todas as classes de GPCRs. Os resultados desta pesquisa fornecem ferramentas poderosas para a descoberta de fármacos e informações biológicas valiosas sobre as características que compõem os ligantes de GPCR.

Palavras-chave: aprendizado de máquinas, receptor acoplado à proteína G, descoberta de fármacos.

Abstract

GPCRs are crucial receptors for many vital physiological processes including control of cell division and proliferation, regulation of ion transport, modulation of neuronal firing, homeostasis, modulation, and modification of cell morphology. They are also involved in many pathological processes, such as in Alzheimer's and Parkinson's disease, cardiovascular disorder, asthma, depression, and diabetes. Given their biological importance, over a third of FDA approved drugs target GPCRs. Nonetheless, GPCRs lead compound development suffers from high attrition rates, with poor *in vivo* efficacy being the primary contributor, resulting in only 7% of all drugs (for other receptors as well) in phase I studies being marketed. This thesis focused on the development of machine learning models capable of predicting bioactivity of small molecules when interacting with GPCRs as means to support the discovery of novel leads through ranking compounds on drug discovery investigations, which would enable enriching screening libraries with compounds more likely to be active. The developed models (composing the pdCSM-GPCR tool) rely on deriving a range of molecular signatures from known ligands, associating them to bioactivities, and modelling them as regression problems, making them independent of receptor structural information. Because of this characteristic, the same approach can be used for any GPCRs which already had been screened for ligands, and also other important targets, including kinases, and ligand-gated ion channels. Our models make up the most comprehensive computational resource for prediction of GPCR bioactivity to date, including support for drug development for orphan GPCRs. Our approach achieved Pearson's correlations of up to 0.89, across 10-fold cross-validation and blind tests. We significantly outperformed previous methods. pdCSM-GPCR was made freely available via a user-friendly web server at http://biosig.unimelb.edu.au/pdcsm_gpcr. We also investigated the properties of small molecules with high affinity for GPCRs in order to identify molecular determinants of recognition. Overall, potent ligands possess nitrogen-containing fragments and aromatic rings, features common in ligands across all classes of GPCRs. The outcomes of this research provide powerful tools for GPCR drug discovery and valuable biological insights into the characteristics that make up GPCR ligands.

Keywords: machine learning, G protein-coupled receptors, drug discovery

List of Figures

List of Tables

List of abbreviations and acronyms

7TM	Seven-transmembrane
AC	Adenylyl cyclase
ADMET	Absorption, Distribution, Metabolism, Excretion and Toxicity
CID	Compound ID number
CSM	Cutoff scanning matrix
ECFP	Extended-Connectivity Fingerprints
ECL	Extracellular
FDA	Food and Drug Administration
GABA	Gamma-aminobutyric acid
GDP	Guanosine diphosphate
GEFs	Guanine nucleotide exchange factors
GIRK	G protein-regulated inward-rectifier K+ channels
GPCRs	G protein-coupled receptors
GRKs	G protein-coupled receptor kinases
GTP	Guanosine triphosphate
ICL	Intracellular
Log P	Lipophilicity
mGluR	Metabotropic glutamate receptors
ML	Machine Learning
MoSS	Molecular Substructure Miner
NAMs	Negative allosteric modulators
PAMs	positive allosteric modulators
PDB	Protein data bank

PI-PLC	Phosphoinositide-specific phospholipase C
RMSE	Root Mean Square Error
SBDD	Structure-Based Drug Discovery
SMARTS	SMILES arbitrary target specification
SMILES	Simplified Molecular Input Line Entry Specification
TM	Transmembrane

Contents

1 Introduction

1.1 G protein-coupled receptors

Every human cell, in order to maintain its inner machinery homeostasis, needs to be able to receive information from the outside and from other cells. This information can be hormone levels, odour molecules, neurotransmitters, among others. This wide transmembrane communication system is supported by G protein-coupled receptors (GPCRs), enzyme-linked hormones, and ligand-gated ion channels. GPCRs comprehend the largest family of transmembrane receptors. They are seven-transmembrane domain proteins located in the plasma membrane and are pivotal as signal transducers for many essential physiological processes such as control of cell division/proliferation, modulation of neuronal firing, homeostasis, regulation of ion transport across the plasma membrane, and modification of cell morphology (NEW; WONG, 2007). Other common names for these receptors include: serpentine receptors, seven-(pass)-transmembrane domain receptors, heptahelical receptors, and G protein-linked receptors. These receptors are responsible for responding to approximately two-thirds of hormones and neurotransmitters (FOSTER et al., 2019) and account for 4% of human genes (KOOISTRA et al., 2020). Such as their importance in biology that they are conserved from excavates to animals and constitute one of the major eukaryotic signalling pathways (MENDOZA; SEBÉ-PEDRÓS; RUIZ-TRILLO, 2014). Related to their key role in physiology, it is expected that they are associated with many human diseases, including Alzheimer's and Parkinson's, cardiovascular diseases, asthma, strokes, diabetes insipidus (SALON; LODOWSKI; PALCZEWSKI, 2011). As a result of the previous statements, GPCRs are largely studied as drug targets (ZHANG; XIE, 2012). It is estimated that drugs targeting GPCRs accounts for more than one third of all authorised drugs by the United States Food and Drug Administration (FDA) (HAUSER et al., 2017).

When GPCRs bind to natural ligands or drugs, their transmembrane (TM) disposition permit the transformation of extracellular messages into intracellular responses (SALON; LODOWSKI; PALCZEWSKI, 2011), allowing them to selectively bind to many types of ligands, ranging from light-sensitive compounds, ions, amino acids, peptides, neurotransmitters, hormones, pheromones and odorants, and send signals from the outside of the cell to its intracellular side. This passage of information is accomplished through regulation of coupling and decoupling of heterotrimeric G proteins or arrestins. The message is magnified and regulates cell physiology.

These receptors have been studied for more than 100 years. John Newport Langley, a British physiologist, was the first to mention them in 1905, in a classic paper, in which he talked about "receptive substance". He proposed that in all cells at least two components are to be differentiated, a "chief substance" which is related to a function and a "receptive substance" that is capable of regulating cell behaviour (MAEHLE, 2004). Sir Henry Dale, who was Langley's

student, kept working with GPCRs and suggested that variation of the binding affinities of adrenaline to "receptive substance" could cause the difference in how cells were affected. Five years later an American pharmacologist, Raymond Perry Ahlquist, put forward the idea of the existence of two types of adrenaline-receptors, alfa and beta (MAEHLE, 2009). These pioneering studies provided perspectives to differentiate GPCRs.

GPCRs form a multigene family consisting of around 800 genes in humans (NIIMURA, 2009). Despite the common transmembrane topology, GPCR tertiary structures are also very diverse, they differ in the sizing of cytoplasmic loops, extracellular amino-terminal tails, and carboxy-terminal tails. Considering these structural differences, GPCRs are grouped into five families: rhodopsin-like (class A), secretin-receptor-like (class B1), metabotropic glutamate receptor (class C), adhesion receptor (class B2), frizzled/taste2 receptor (class F) (FREDRIKSSON et al., 2003) (see Figure 1). Among these classes it is important to mention that there are some receptors called orphan GPCRs, whose endogenous ligands remain unidentified, leaving their natural functions in doubt, and can be a great source of drug targets

Class A comprehends most of the GPCRs. There are 719 members in this family. It is divided into subfamilies such as: melatonin, nucleotide, aminergic, peptide, protein, lipid, steroid, alicarboxylic acid and sensory (YANG et al., 2021a). Most drug discovery research involving GPCRs focus on this class. Currently, there are over 500 drugs targeting it. These aforementioned drugs are used for cancer treatment, allergies, migraine, pulmonary diseases, hypertension, depression, cardiovascular diseases, glaucoma, Parkinson's disease, schizophrenia and as analgesics (YANG et al., 2021a).

Class B is split up into two subfamilies: secretin (B1) and adhesion (B2). While the former contains 15, the latter contains 33 members (ALEXANDER et al., 2019). Secretin subfamily members interact with vasoactive intestinal peptide, calcitonin gene-related peptide, corticotropin-releasing factor, glucagon, pituitary adenylate cyclase-activating peptide, parathyroid peptide hormone, growth hormone-releasing hormone, and glucagon-like peptides (ALEXANDER et al., 2019). Adhesion subfamily are differentiated from other GPCRs due to their functions in cell adhesion and migration (BHUDIA et al., 2020). In this class, glucagon family peptides receptors, calcitonin gene-related peptide, parathyroid peptide hormone, corticotropin-releasing factor, vasoactive intestinal peptide, growth hormone-releasing hormone, and pituitary adenylate cyclase-activating peptide, are the major therapeutic targets. They were developed as treatments for obesity, type 2 diabetes mellitus, osteoporosis, migraine, depression, and anxiety (ALEXANDER et al., 2019).

Class C, which are also called glutamate receptors, have their activation linked to indirect metabotropic processes. A distinctive feature of this class is that these receptors are obligated constitutive dimers for receptor activation (PIN et al., 2005). This class contains 22 receptors, which are categorized into 5 subfamilies: calcium-sensing receptor (1 receptor), gamma-aminobutyric acid (GABA) type B receptors (2 receptors), taste 1 receptors (TS1R1–3,

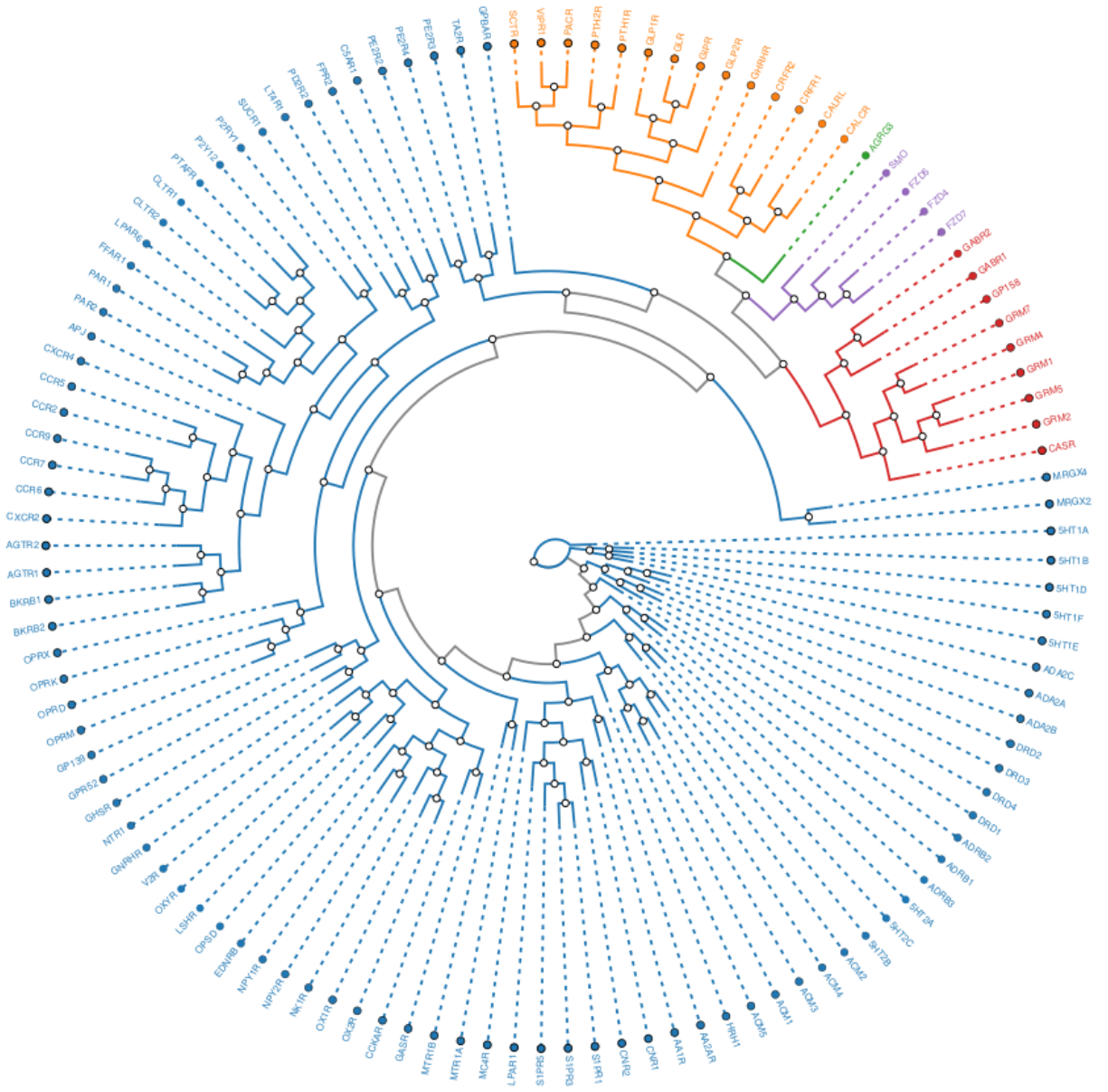


Figure 1 – GPCR phylogenetic tree with all solved GPCR structures available at GPCRdb (2021). Class A are represented by the colour blue and are predominant, class B1 by orange, class B2 by green, class C by red and class F by purple. The tree was calculated using 10 replicas of Bootstrapping, and the distance calculation method was Neighbour-joining. Sequence segment selection: Structurally conserved (generic) positions. This phylogenetic tree was built using GPCRdb website

3 receptors), metabotropic glutamate receptors (mGluR1–8, 8 receptors) and 8 orphan GPCRs (which are proteins that that have not yet been thoroughly characterised or classified) (NISWENDER; CONN, 2010). Up to now, 16 drugs have been approved by the FDA targeting 8 class C GPCRs and are used for cancer treatment, schizophrenia, depression, and movement disorders (ALEXANDER et al., 2019).

The smaller class of all is F, which includes only 10 different types of Frizzled and one smoothened receptor. The former is involved in the Wnt signalling pathway, and the latter in

the Hedgehog signalling pathway. The smoothed receptor is a validated target for antineoplastic agents (RUAT et al., 2014).

1.1.1 Structure of G protein-coupled receptors

As previously mentioned, GPCRs represent the largest protein family in the human genome (CHEN et al., 2019), with great diversity in terms of their amino acids sequences. Despite that, all of them possess common structural features, including a domain comprising seven-transmembrane (7TM) helices linked by three extracellular (ECL) and three intracellular (ICL) loops (see Figure 2 (CIANCETTA et al., 2015)). The TM component exhibits great identity between all proteins of this family. The most variable portions of GPCRs are the carboxyl terminus, the intracellular loop spanning TM5 and TM6, and the amino terminus.

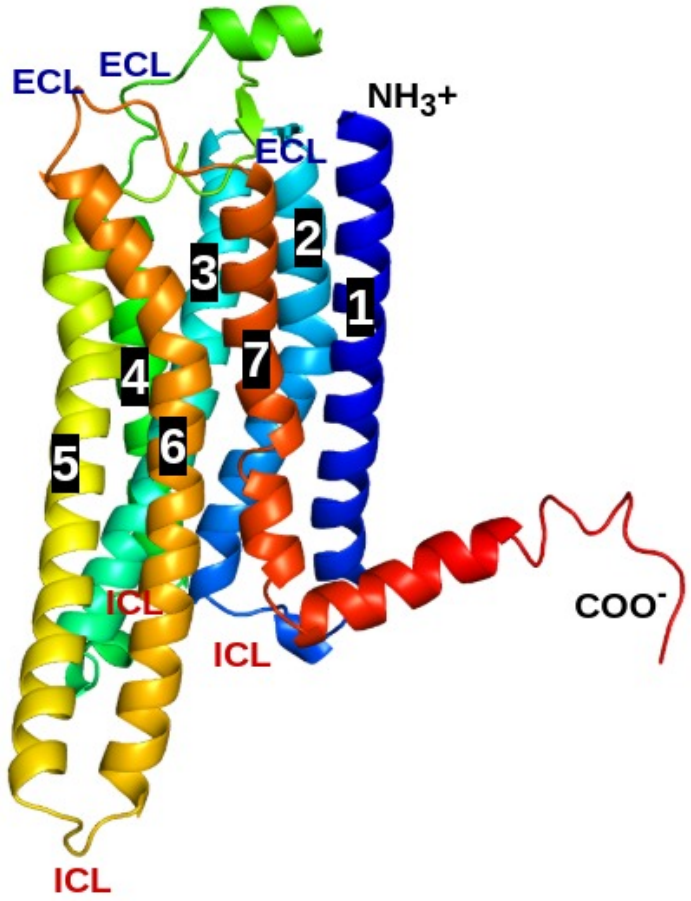


Figure 2 – General GPCR structure. The seven-transmembrane Adenosine receptor A2 structure, PDB template 2YDO. These membrane proteins comprise seven transmembrane helices connected by ICL and ECL loops. Their amino part faces outside the cell and their carboxyl part the inside of the cell (CIANCETTA et al., 2015).

Besides the high structural similarity on 7TM regions among all GPCRs, there are a series of conserved sequence motifs (micro-switches) typical for each class (see Figure 3). For example, in class A, it is found in TM helix 3, a highly conserved "DRY" motif, located at the bottom of the 7TM. This motif springs an intra-helical salt bridge between D/E$^{3.49}$ and R$^{3.50}$ (the

two numbers in superscript represent residue numbering by (BALLESTEROS; WEINSTEIN, 1995), where the first denotes the helix, 1–7, and the second the residue position relative to the most conserved residue, defined as number 50 (arbitrary). In the first case, 3.49 denotes a residue located in TM3, one residue before the most conserved residue), and it is known to stabilise receptors in an inactive state. The binding of agonists triggers a rotameric switch of $W^{6.48}$, in an also highly conserved motif known as "CWxP". Another common microswitch is the "NPxxY". When the receptor is activated, the residue $Y^{7.53}$ of this motif alters its rotamer conformation and points toward TM3, making new contact formation between $Y^{7.53}$ and residues in TM3 (ZHOU et al., 2019). Changes in the hydrogen bonding network in this microswitch indicates a possible mechanism of enhanced thermal stability (WHITE et al., 2018). At last, in class A, it is worth mentioning the motifs PIF and the Na+ pocket residues (an allosteric site where sodium ions bind, an example is the residue $Asp^{2.50}$ (SELENT et al., 2010)), which suffer rearrangements during receptor activation. All these cited conserved motifs are critical for the activation of class A GPCRs (YANG et al., 2021b).

In Class B receptors, the binding of a peptide causes destabilisation of the TM6 helix, and hence it initiates a sharp kink shaping at the conserved motif $P^{6.47b}xxG^{6.50b}$ (YANG et al., 2021b). Class C are distinguished by a large extracellular domain that forms an obligate dimer. It also contains conserved motifs, the motif 'F/Y/HxPKxY' on TM7 and 'Fx WxP' on TM6 (DORÉ et al., 2014). Class F GPCRs possess the conserved 'KTxxxW' motif. These common motifs, during activation, go through considerable conformational change (BERTALOVITZ et al., 2016).

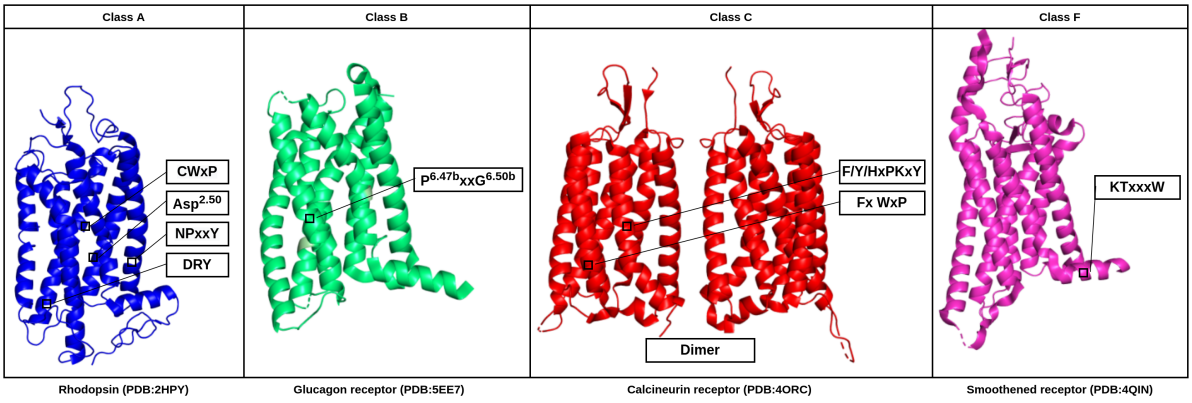


Figure 3 – GPCR structures. Crystal structures of representative GPCRs from classes A, B, C, and F with their respective conserved sequence motifs (micro-switches)(YANG et al., 2021b; DORÉ et al., 2014; BERTALOVITZ et al., 2016).

In 2000, the first crystal structure of a GPCR was elucidated. It was a rhodopsin purified from bovine eyes, in an inactive state (PALCZEWSKI et al., 2000), a naturally abundant and highly stable GPCR, characteristics that enabled this discovery. Nonetheless, usually it is challenging to obtain GPCR diffraction-quality crystals for high-resolution structure determination (KOBILKA; DEUPI, 2007), mostly due to their intrinsic conformational complexity and low

expression yield. In order to obtain high-resolution crystals, engineering is often required to minimise conformational heterogeneity and maximise crystal contacts and stability (CHUN et al., 2012; MAGNANI et al., 2008; SCOTT et al., 2013; BILL et al., 2011; THAL et al., 2018). In the last decades, several methods have been developed in this sense, which include recombinant over expression, purification strategies (ERREY; FIEZ-VANDAL, 2020), crystallisation platforms (PARKER; NEWSTEAD, 2012) and detergent studies (LEE et al., 2020). During 2007, the first human GPCR structure was elucidated, the β2-adrenergic receptor bound to an antagonist (CHEREZOV et al., 2007). This breakthrough was reached thanks to innovative methods, which include incorporation of a soluble fusion partner and lipidic cubic phase crystallisation and enhancements in protein expression and purification.

Recent advances in cryo-EM is now turning the wind in favour of this technology and, in 2019, the number of membrane protein structures solved by cryo-EM became higher than by X-ray crystallography. This happened mainly because the former does not require crystallisation and microgram amounts of protein and the latter, besides crystallisation, requires milligram amounts. Another point is that cryo-EM can tolerate certain degree of sample impurity and heterogeneity and can also reach resolutions below 3Å (similar to resolutions obtained through X-ray crystallography) (GARCÍA-NAFRÍA; TATE, 2021).

1.1.1.1 Computational prediction of GPCR structures

In the lack of experimentally determined structures, computational tools are an alternative for elucidating GPCRs structures (ZHANG et al., 2015; WORTH et al., 2011; LAUNAY et al., 2012; SANDAL et al., 2013; ESGUERRA et al., 2016; COSTANZI et al., 2016). One example of such strategy is homology modelling. For employing this method, it is necessary the existence of another protein, a close homologue (called template), with an experimentally determined structure (WEBB; SALI, 2016). The unknown structure is predicted taking into account the sequence identity between the target and the template and the structure of the latter. This process is based on the fact that protein structures are more conserved than protein sequences amongst homologue proteins (CHOTHIA; LESK, 1986). One problem with these predictions for GPCRs is that although GPCRs shares the seven TM architecture, the loops, on the other hand, exhibit huge structural diversity and low sequence conservation, especially in the ECL2 portion (KATRITCH; CHEREZOV; STEVENS, 2012; WOOLLEY; CONNER, 2017). This lack of sequence identity information hinders accurate structure prediction of some GPCR loops (BUSATO; GIORGETTI, 2016). In this context, some studies tried using ab initio methods (FISER; DO; ŠALI, 2000; JACOBSON et al., 2004; SPASSOV; FLOOK; YAN, 2008). These approaches search for the loop conformational space on the energy landscape without using of known structures data (WON et al., 2018). Nevertheless, ab initio can be deployed only for small segments (100 residues), because it demands vast computational resources (LEE; FREDDOLINO; ZHANG, 2017).

Latterly, machine learning (ML) is revolutionising structure prediction. Examples of

new ML tools are called AlphaFold2 (JUMPER et al., 2021) and RoseTTAFold (BAEK et al., 2021). These ML models are able to generate accurate structure prediction for basically almost any sequence, even when homologue proteins are not available. The aforementioned methods were built using neural network models that have learned how to infer inter-residue interactions and protein structure using knowledge abstracted from various known experimental structures. Nevertheless, it is important to note, that even for these novel methods, there are some incongruities in the prediction for GPCRs. One example is ECL2 (regarded to play a critical role in ligand recognition) is often indicated as a low or very low confidence region on the predicted structure (NICOLI et al., 2022). Another limitation regards to representing structural dynamics that can lead to multiple conformations. GPCRs exists in multiple conformational states, essential for their signalling roles. Even so, these prediction methods are trained to predict a solo, native state for a given sequence. For GPCRs, the state predicted is usually the inactive. This probably is a reflection of overrepresentativeness of inactive states in structural data banks (HEO; FEIG, 2021).

1.1.2 Mechanics of receptor activation

The mechanics of receptor activation, in summary, involve three steps: 1) ligand binding, 2) generation of signalling, and 3) transduction of signalling throughout the cell. In step one, the binding of a ligand induces GPCRs to act as guanine nucleotide exchange factors (GEFs). GEFs are protein domains or proteins that activate monomeric GTPases by inducing the release of guanosine diphosphate (GDP) to permit binding of guanosine triphosphate (GTP) (CHERFILS; ZEGHOUF, 2013). This change triggers step two, in which the ligand-bound GPCRs causes an exchange of GDP to GTP (MCCUDDEN et al., 2005) in the protein G. This exchange causes the dissociation of the Gα subunit from the dimer G$\beta\gamma$ dimer and from the receptor. Both the Gα and the dimer interact with other intracellular components, causing cascades of events and continuation of the transduction signalling, which constitutes step three. After a while, the GTP is hydrolysed to GDP, and it allows the reassociation of its heterotrimeric portion (DIGBY et al., 2006) (Figure 4).

GPCRs not only activate heterotrimeric G proteins, but also promote receptor phosphorylation by G protein-coupled receptor kinases (GRKs) and subsequent binding of beta-arrestins that induces additional signalling cascades (LEFKOWITZ, 2013). Beta-arrestins are also important for the GPCR desensitisation, endocytosis and signalling control (RANKOVIC; BRUST; BOHN, 2016). Besides the mentioned GPCR protein signalling system, there are alternative upstream and downstream molecules, such as RGS and GoLoco which can regulate G protein heterotrimeric signalling. Ric 8 (resistance to inhibitors of cholinesterase 8) can cause activation of GPCR-independent and phosducins, which can also regulate beta/gamma subunits (Figure 4).

The complexity and speciality of GPCR signalling relies on, to a certain extent, the existence of different types of G-protein subunits (HERMANS, 2003). For instance, Gα effectors

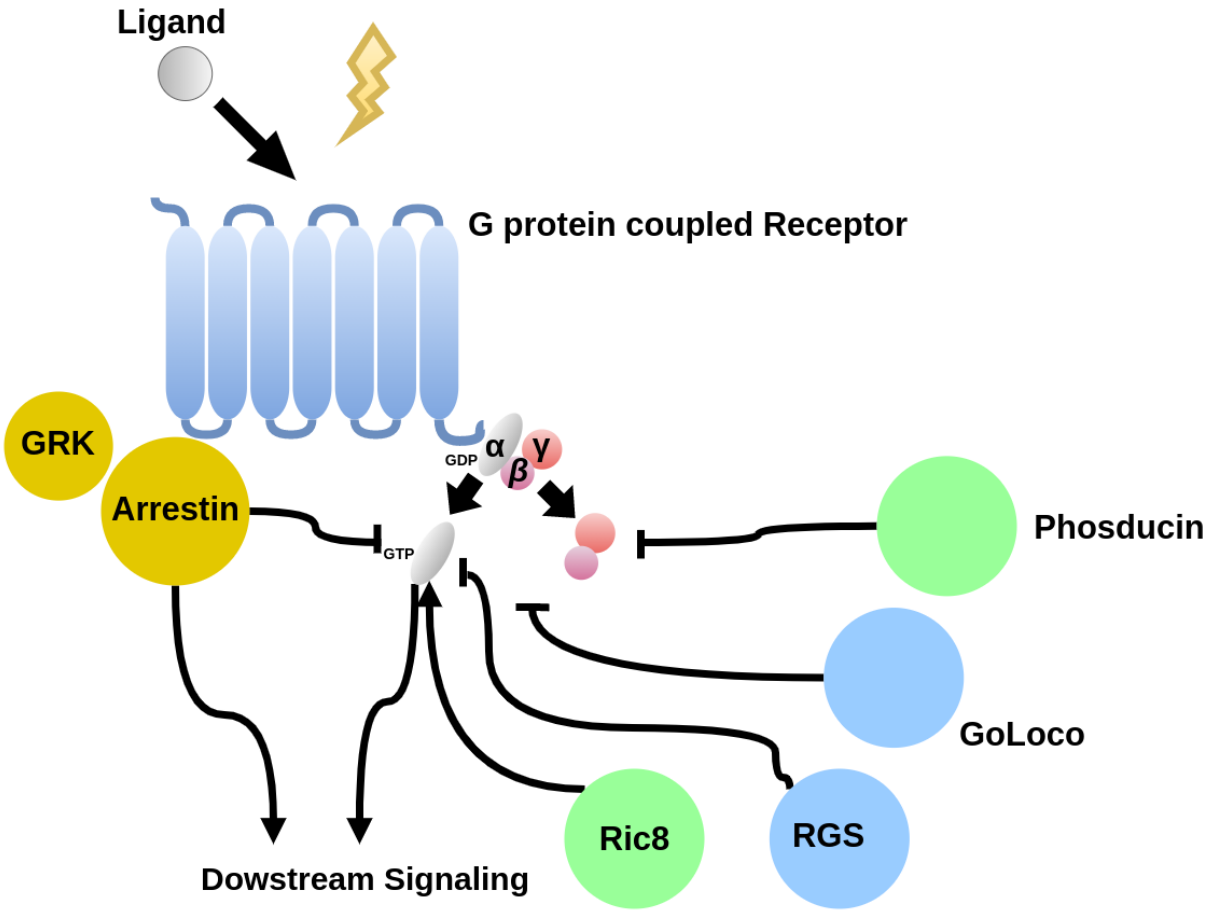


Figure 4 – GPCR mechanism. Schematic representation of the GPCR signalling pathway (adapted from de Mendoza et al., 20). During inactivation, GPCRs (in blue) interact with a G alpha (gray oval shape) bound to GDP, G beta-gamma (purple and red circles). Upon receptor stimulation by agonist, an exchange of the GDP bound to the G alpha for a GTP, causes G alpha to dissociate from the receptor and G beta-gamma, which causes G alpha activation. G alpha then goes on to trigger other molecules in the cell. Besides the mentioned GPCR protein signalling system, there are alternative upstream and downstream molecules. For instance, RGS and GoLoco can regulate G protein heterotrimeric signalling. Ric 8 (resistance to inhibitors of cholinesterase 8) can cause activation GPCR-independent, beta/gamma subunits are also regulated via phosducins. GPCRs can perform downstream signalling independently of G proteins by GRKs and Arrestins (MENDOZA; SEBÉ-PEDRÓS; RUIZ-TRILLO, 2014).

can be classified in classes, Gαs, when it stimulates adenylyl cyclase (AC), Gαi, when inhibits AC and thus opposes the action of Gαs, Gαgust and Gαolf when it acts as tastant and odorant receptors, respectively. The Gα involved in vision is termed as Gαt and regulates a cyclic GMP-gated Na+/Ca2+ channel through its effector. There are G-protein subunits termed as Gαq class (this class also includes, Gα11, Gα14 and Gα16) which activates phosphoinositide-specific phospholipase C (PI-PLC) isozymes and also Gα12/13, which can regulate the small G-protein RhoA (MCCUDDEN et al., 2005). On the other side, G$\beta\gamma$ dimer that was once thought to only help coupling of G$\alpha\beta\gamma$ heterotrimers to GPCRs and act as a Gα inhibitor, now also is known to interact with numerous effectors, after dissociation of Gα-GTP. The first G$\beta\gamma$ effectors identified were the G-protein-regulated inward-rectifier K+ channels (GIRK). Afterwards, it was discovered that G$\beta\gamma$ subunits can also regulate many kinases and small G-proteins (MCCUDDEN et al., 2005).

Another interesting fact regarding G proteins is that GPCRs can couple with discrete G-proteins, which one leading to the activation of multiple intracellular effectors. This mechanism increases the complexity and specificity of GPCR signalling. Just as an example, a study found that a single receptor can at the same time activate members of four classes of G-proteins (Gs, Gi/o, Gq/11, and G12) (LAUGWITZ et al., 1996) (see Table 1).

Table 1 – G-proteins and effectors, ↑=increase, ↓=decrease, (Hermans [2003]).

Subunit	Family	Main subtypes	Primary effector
α	αs	Gαs, $G\alpha olf$	Adenylate cyclase ↑
	$\alpha i/o$	G$\alpha i-1$, $G\alpha i-2$, $G\alpha i-3$	Adenylate cyclase ↓
		GαoA, $G\alpha oB$	K+ channels ↑
		G$\alpha t1$, $G\alpha t2$	Ca2+ channels ↓
		Gαz	Cyclic GMP phosphodiesterase ↑
	$\alpha q/11$	Gαq, $G\alpha 11$, $G\alpha 14$, $G\alpha 15$, $G\alpha 16$	Phospholipase C ↑
β	$\alpha 12$	G$\alpha 12$, $G\alpha 13$	Adenylate cyclase ↑/↓
	$\beta 1\check{}5$		Phospholipases ↑
			Phosphatidylinositol 3-kinase ↑
γ	$\gamma 1\check{}11$		Protein kinase C ↑
			Protein kinase D ↑
			GPCR kinases ↑
			Ca2+, K+ (and Na+) channels

1.1.3 Binding sites

In the last years, GPCR structure elucidation has greatly demanded stabilisation using ligands (ZHANG et al., 2015). To date, the binding of small molecules to elucidate structures covers all activity types, from agonism to antagonism, although the latter is most frequently employed, mainly because the inactive state of GPCRs tends to be more stable and easier to crystallise (CONGREVE et al., 2020). This ligand-data availability subsidised the discovery

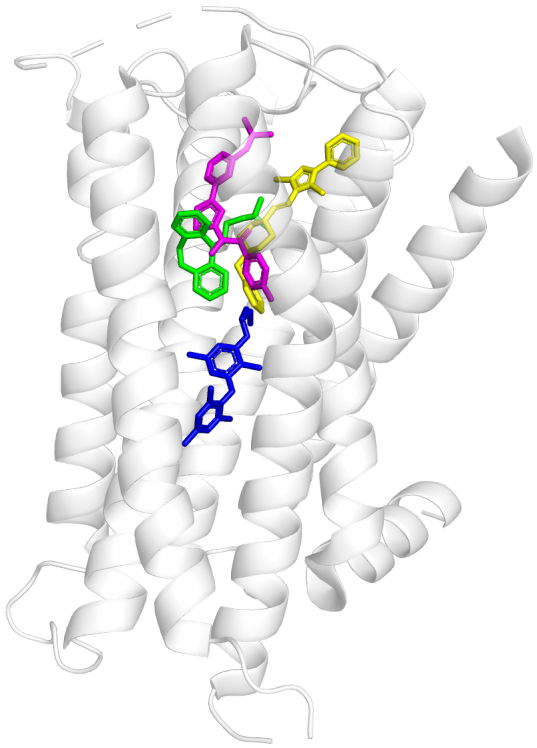


Figure 5 – Depth of ligand binding in the transmembrane pocket for the GPCR classes A, B, C and F. The histamine H receptor (class A: CXCR4-vMIP-II, green ligand, PDB: 3RZE) is displayed as transparent white cartoon, and was used for the superposition with other structure complexes, class B: CRF1R-CP-376395 (blue, PDB: 4K5Y), class C: mGlu1-FITM (pink, PDB: 4OR2) and class F: smoothened receptor-SANT-1 (yellow, PDB: 4N4W) (adapted from (MUNK et al., 2016).

of many ligand binding sites. Most of the ligands have traditionally been acknowledged to interact with orthosteric sites in the upper region within the transmembrane domains. However, there are variations involving the depth of penetration into the transmembrane pocket (Figure 5), shows one ligand for each class). For class A, the most superficial ligand found is in the histamine receptor (SHIMAMURA et al., 2011) and the deepest in the chemokine CXCR4 receptor (WU et al., 2010). In Class B, CRF1 receptor has the deepest ligand binding site found co-crystallized with ligand CP376395 (HOLLENSTEIN et al., 2013). Class C mGlu1 negative allosteric modulator FITM largely overlaps with class A ligands (WU et al., 2014). Class F receptor, smoothened (SMO), has been crystallised with multiple ligands, some very close to the extracellular surface (WANG et al., 2013), and others covering much deeper areas in the transmembrane region (WANG et al., 2014).

1.1.4 G protein-coupled receptor ligands

Considering GPCRs key role in many biological processes and diseases, and their distribution across nearly all organs and tissues (HAUSER et al., 2017), their great importance in drug development is inherent (ZHANG; XIE, 2012). Nevertheless, there are only 134 GPCRs

that are a target of current approved drugs (SRIRAM; INSEL, 2018). This corresponds to approximately 16% of the GPCRs universe (800 in humans) (including olfactory receptors) (SRIRAM; INSEL, 2018). The remainder of GPCRs could be involved in fundamental parts of the immune system, genetic, neurological and metabolic disorders (HAUSER et al., 2017), and may prove to be interesting novel drug targets (DIAZ; ANGELLOZ-NICOUD; PIHAN, 2017). Following this assumption, it is evident that there is much space for improvement in this domain, and elucidation of new GPCRs structures and discovery of new GPCRs ligands are of paramount importance.

According to Mason et al. 2012 (MASON et al., 2012), the evolution of many GPCR receptors culminated in the binding of low molecular weight, fragment-like ligands, such as acetylcholine and dopamine. It is also common that drugs that interact with this receptor are normally analogues of natural ligands. They estimated that 70% of the drugs that target GPCRs are natural ligand-related. Therefore, the discovery of new ligands can be based on molecules that are known to bind these receptors. It is also mentioned by these researchers that 15% of GPCRs drugs already on the market were targeted serendipitously, which means that they were discovered using phenotypic assays and without the knowledge of their mechanism of action. Later, many mechanisms of action belonging to these drugs were elucidated and these pieces of information are of great interest for the optimisation of identified ligands and discovery of new ones. Some examples of important drugs targeting GPCRs include: opioid analgesics, antihistamines, anticholinergics, typical and atypical antipsychotics antimigraine drugs, β2-agonists for asthma and antihypertensives (WACKER; STEVENS; ROTH, 2017).

GPCR modulators can be divided into full agonists, partial agonists, antagonists, and inverse agonists. Full agonists can induce maximal GPCR activity, partial agonists, on the other hand, cannot cause maximum activation of receptors and can act as an antagonist if competing by the binding site with a full agonist. Antagonists are agonist blockers and can be classified as neutral antagonists and inverse agonists. The former binds to GPCRs but does not affect the receptor's constitutive activity, the latter induces a pharmacological response opposite to that of the agonist, by suppressing spontaneous receptor signalling (when present). All these modulators interact with orthosteric binding sites of GPCRs (WACKER; STEVENS; ROTH, 2017). While, as already mentioned, TM regions in GPCRs are very conserved, ECL and ICL regions possess remarkable diversity. The ECL region, more specifically in ECL2 is regarded to play a critical role in ligand recognition, access, and selectivity (DROR et al., 2011; KRUSE et al., 2012; ZHANG et al., 2015). This region connects TM4 and TM5 and contains a highly conserved cysteine which forms a disulphide bridge with TM3. As mentioned, ECL2 is contiguous with TM5, whose motion is determinant for GPCR activation. (NICOLI et al., 2022) analysed the number of covalent and non-covalent contacts between ECL2 and TM5 for all structures solved in active and inactive conformations. They found that, in most cases, the number of contacts between these two regions decreases in the inactive conformation. Also, according to them, if this observation is confirmed when more structures are available, can be of great importance for

implementing GPCR structure-based drug design workflows.

Besides looking for orthosteric site ligands when studying GPCRs, allosteric modulators provide a great advantage, as they can confer greater selectivity, because these molecules tend to be less well conserved between related receptors compared to orthosteric sites, and can even distinguish between closely related receptor subtypes (FOSTER et al., 2019), thus inducing less side effects. This tendency of less conservation, is due to the fact that they have not confronted the same evolutionary pressure as orthosteric sites to fit an endogenous ligand, so that drug specificity is more feasible to be achieved (CHRISTOPOULOS et al., 2004). When these modulators bind GPCRs they cause changes in receptor conformation, so that it interferes with their interaction with orthosteric ligands. These modulators can be divided into several distinct types: negative allosteric modulators (NAMs), which act inhibiting receptor activation via negative cooperativity reducing the affinity and/or efficacy of orthosteric agonists; partial NAM, which does not completely block receptor activation; positive allosteric modulators (PAMs), which increases agonist affinity and, consequently, increases the potency and/or efficacy of orthosteric agonists; and Ago-PAM which can work as PAMs or can induce receptor activation even in the absence of orthosteric ligands (BASITH et al., 2018; FOSTER; CONN, 2017). A classical example of positive allosteric modulators of GABA receptors is Benzodiazepines, an effective and safe approach to the treatment of anxiety and sleep disorders (MöHLER; FRITSCHY; RUDOLPH, 2002). Allosteric modulators of GPCRs are now being evaluated as potential drug candidates for Parkinson's disease, schizophrenia, Alzheimer's disease and dystonia (CONN; CHRISTOPOULOS; LINDSLEY, 2009; KRUSE et al., 2014).

1.1.5 Targeting G protein-coupled receptors

GPCRs have enormous physiological and biomedical importance, being the primary site of action of approximately 34% of prescribed drugs (HAUSER et al., 2017). As already mentioned, the current drugs target only a few GPCRs and thus there is a unique opportunity to design new treatment exploring these receptors.

One key factor that complicates drug design for these proteins is that GPCRs exist, according to molecular dynamics simulations, in several "intermediate" conformational states between the crystallographic active and inactive states. In molecular dynamics states, there is a notable intermediate where TM7 adopts an inactive conformation while TM6 remains in an active state (LATORRACA; VENKATAKRISHNAN; DROR, 2016). Besides this noteworthy intermediate state, there are many others and the transition between active and inactive conformational states can take different routes to be achieved (LATORRACA; VENKATAKRISHNAN; DROR, 2016). Further indications of the diversity of conformational states are found in crystal structures of different GPCRs. Even when these receptors are bound to their agonists, a great part of GPCRs crystals are not in a fully active conformation capable of signalling (LEBON; WARNE; TATE, 2012). The importance of this conformational diversity is related to the biassed

signalling observed in GPCRs. On the active state, GPCRs typically couple to both G protein and also arrestins, and both downstream signalling pathways are activated. Nonetheless, some ligands favours one downstream pathway in detriment to the other and, as a result, we have some small molecules with a molecular response based on G protein signalling and others on arrestin signalling. This biassed signalling is related to these different ligands selecting not just one type of active and one type of inactive conformational state but capable of selecting among multiple conformational states with different abilities to couple to different downstream partners (STRACHAN et al., 2014; VIOLIN et al., 2014). The study of biassed agonism can render safer GPCR drugs due to their potential to exclusively activate desired signalling pathways instead of activating a pathway which leads to side effects (NAGI; ONARAN, 2021; BOCK; BERMUDEZ, 2021). Besides, it is important to mention that being able to bind to a GPCR does not necessarily mean that the ligand will induce the desired effect, given that some disease treatments involve targeting the inactive state (such as beta blockers (KOBILKA, 2011)) and others the active states (salbutamol, for example (BHATTACHARYA et al., 2008).

Given the conformational complexity of ligand-activated GPCRs, and also that most of their surface is buried inside the membrane, it is not surprising that many structural modifications are required in order to obtain high quality three-dimensional crystal structures. Recently, the development of ingenious engineering techniques, such as stabilisation of TM5/TM6 region through T4-lysozyme insertion, stabilisation of TM5/TM6 region through binding to fragment antigen-binding region (Fab region)(ROSENBAUM; RASMUSSEN; KOBILKA, 2009; MILIC; VEPRINTSEV, 2015), mutations to increase thermal stability and functional expression, have rendered crystallographic structures for a broad range of GPCRs. These structural discoveries coupled with the evolution of computational methods (molecular dynamics, integrative modeling and machine learning (ZHU et al., 2021)) led to the development of high-quality models which are now freely available in dedicated repositories such as the GPCRdb (PÁNDY-SZEKERES et al., 2017) and GPCR-EXP. For instance, (PÁNDY-SZEKERES et al., 2017) and (LANGMEAD et al., 2012) demonstrated effective lead identification targeting adenosine A2A receptor applying Structure-Based Drug Discovery (SBDD) and disclosed candidates for possible treatment of Parkinson's disease using biophysical mapping and co-crystallised receptors with ligands. (LANGMEAD et al., 2012) carried out an *in silico* screening of, over half a million compounds, using the homology model of the β1 adrenergic receptor (based on the crystal structure of the turkey β1 adrenergic receptor complexed with cyanopindolol). The outcome of this study was 20 confirmed hits *in vitro*. (CHRISTOPHER et al., 2015) did a fragment screening of a thermostabilised mGlu5 receptor and, after this procedure, he used a SBDD approach to optimise potential candidates and developed a high potent series of negative allosteric modulators for this metabotropic GPCR. Besides these studies, it is important to mention some reviews that focused on identification of ligands for orphan GPCRs. (NGO et al., 2016) evaluated methods used to establish the appropriate signalling assays to test orphan receptor activity; they also studied cases of structure-based methods for targeting orphan GPCRs. In 2015 Huang et al. (HUANG et al.,

2015), used a yeast-based screening against the understudied orphan GPCR, GPR68, and also SBDD and identified the benzodiazepine drug lorazepam as a non-selective GPR68 positive allosteric modulator. Jiménez-Rosés et al., 2021 (JIMÉNEZ-ROSÉS et al., 2021) assembled a database of around 2,700 known β 2AR agonists and antagonist ligands and computationally docked them to multiple experimentally determined β 2AR structures. For each one of 75,000 docking poses, they identified specific interactions and correlated them with agonist or antagonist activity. Afterwards, they were capable of developing machine learning (ML)-based predictors of agonist/antagonist activity with up to 90% accuracy (JIMÉNEZ-ROSÉS et al., 2021), demonstrating that it is possible to use ML and current data in GPCRs ligand discovery.

It is important to state that the absence of structures for numerous GPCRs has restrained the ability to apply rational structure-based drug development for some receptors (HEIFETZ et al., 2015). An alternative when there is lack of structural information on specific GPCRs are ligand-based approaches combined to machine learning (JABEEN; RANGANATHAN, 2019). This strategy is supported by the existence of large datasets of molecules known to bind GPCRs, (PÁNDY-SZEKERES et al., 2017), Drugbank (WISHART et al., 2017), PubChem (KIM et al., 2018), ZINC (STERLING; IRWIN, 2015), CHEMBL (GAULTON et al., 2016) and BINDINGDB (GILSON et al., 2015). These datasets are used to derive pharmacophore models based on the physicochemical characteristics and atom patterns of known ligands without prior knowledge of the protein structure. The derived pharmacophore models are then used to predict new molecular entities that can interact with the target. These relevant endeavors, however, are usually limited to one receptor type. Some examples of these efforts includes all sorts of GPCRs, from Cannabinoid receptor to Olfactory receptors. In 2016 three interesting studies were published involving "Three-Dimensional Biologically Relevant Spectrum" (a three-dimensional similarity array based on shape overlap with known ligands and property similarity) as descriptors for development of predictiors: one for a Cannabinoid receptor (HU et al., 2016), one for an Adenosine receptor (HE et al., 2016) and also one for a Dopamine receptor (KUANG et al., 2016). These works demonstrate that it is possible to effectively extract molecular features that describe GPCR ligands. Other two studies also successfully developed predictors, but for Serotonin receptors. They applied different ligand descriptors from the previous methods: one involving 5-HT7R and 5-HT1AR (KURCZAB et al., 2016), used molecular and structural fingerprints and, the other involving selectivity prediction of 5-HT2BR versus 5-HT1BR (RATAJ et al., 2018), using Neighbouring Substructures Fingerprint. Throughout these two studies, it was possible to identify specific binding interactions between ligand and receptors. Contemplating Olfactory receptors, (BUSHDID et al., 2018) screened ligands using 4,884 chemical descriptors. According to their findings, a support vector machine algorithm accurately predicted the activity of compounds, a fact that was confirmed during *in vitro* experiments. Using Extended-Connectivity Fingerprints (ECFP) and deep neural networks (KOUTSOUKAS et al., 2017) developed predictors of ligands for Dopamine D4 receptor, among other classes of proteins and achieved good classification performance. These case examples

demonstrated that a range of molecular properties and fingerprints can be exploited to effectively identify GPCR ligands.

In the last years, a couple of studies have endeavored to produce general workflows to support ligand discovery for multiple GPCRs classes (WU et al., 2018). Of note, Wu2018 used weighted deep learning and random forest to develop the WDL-RF method. This tool comprising predictions for 26 types of GPCRs (classes A, B, C, and F). In 2019, they launched an iteration of their method (SED) (WU et al., 2019). It couples long ECFPs with deep neural network training using a data set of 16 types of GPCRs (covering classes A, B, C and F). Recently, (SAKAI et al., 2021) used graph convolutional neural networks for encoding ligands features and used this information to design models to predict bioactivity of small molecules against 127 diverse targets, further confirming the effectiveness of graph-based methods for ligand discovery. Table 2 summarises relevant applications of ML for the prediction of GPCR ligands.

Table 2 – ML methods applied for developments of tools to support GPCR ligand discovery (JABEEN; RANGANATHAN, 2019).

GPCR	Dataset	Descriptor calculation	Ref
Cannabinoid receptor	ChEMBL	BRS-3D	(HU et al., 2016)
Adenosine receptor	ChEMBL	BRS-3D	(HE et al., 2016)
Serotonin receptors: 5-HT7R and 5-HT1AR	ChEMBL	Hashed FP, Klekota-Roth FP, MACCS FP, Structural Interaction Fingerprint profiles	(KURCZAB et al., 2016)
Serotonin receptors: 5-HT1BR and 5-HT2BR	MCule	Klekota–Roth fingerprint (KRFP) substructure keys	(RATAJ et al., 2018)
Olfactory receptors: OR51E1, OR1A1, OR2W1 and MOR256-3	Literature	Dragon software	(BUSHDID et al., 2018)
Metabotropic glutamate receptor: mGluR1	Literature	DISCOVERY STUDIO 3.1	(JANG et al., 2015)
26 different GPCRs	ChEMBL	Fingerprint generation through novel weighted deep learning	(WU et al., 2018)
Dopamine D4 receptor, Cannabinoid CB1 receptor	ChEMBL	Extended-Connectivity Fingerprints (ECFP)	(KOUTSOUKAS et al., 2017)
16 different GPCRs	ChEMBL	Extended-Connectivity Fingerprints (ECFP)	(WU et al., 2019)

It is also important to cite the high attrition rates in drug development, a fact that is not exclusive for GPCRs. It is estimated that about 10%-20% of the molecules in the beginning of the clinical trial reaches market approval. And this fact has been happening in the past few decades. Moreover, cutting down this attrition rates is a key challenge for all pharmaceutical industry (YAMAGUCHI; KANEKO; NARUKAWA, 2021; WARING et al., 2015). Chemioinformatics can be of great support for this task, because it allows transformation of data into knowledge in a faster pace than most of experimental procedures, enabling the making of better decisions in the drug development area, cutting costs and time of the process (ENGEL, 2006).

In the light of the aforementioned advances in the field of cheminformatics, this thesis focused on the investigation of graph-based signatures, known as Cutoff Scanning Matrix (CSM), combined to various molecular properties to model GPCR ligand activity. These combined features were then applied in the development of predictive models capable of inferring bioactivity of small molecules when interacting with GPCRs. CSM signatures were employed in many cases of success for discovery of small molecule interactions (PIRES; ASCHER; BLUNDELL, 2014;

PIRES; ASCHER, 2016b; PIRES; BLUNDELL; ASCHER, 2015; PIRES et al., 2013; PIRES; ASCHER, 2016a; PIRES; ASCHER, 2017). They use the concept of graph-based signatures, in which the geometry and physicochemical characteristics of the structural environment of a molecule are represented as a network or graph. This characterisation is composed of a series of nodes (which represent atoms) and edges (which describe the distances between the atoms). It is hypothesised that molecules with similar graph-based signatures have similar chemical and biological properties. These signatures were combined to various molecular properties and used as evidence to train and test machine learning models to accurately identify potential GPCR ligands, by developing a computational dedicated platform, pdCSM-GPCR. Our models are capable of quantitatively predicting ligand bioactivity for the most comprehensive set of GPCR types and classes (A, B1, C, and F) to date.

1.2 Justification

GPCRs are essential membrane receptors, involved in a wide range of signalling pathways, along with great involvement in human pathophysiology. GPCRs form the largest human membrane protein family and are the most studied drug targets. Despite the great effort applied to the study of this protein family, structure determination of GPCRs face high failure rates, since the receptors are very unstable and have intrinsic plasticity. Due to that, there are many GPCRs without structure elucidated, including many orphans receptors that can be involved in key biological aspects and disorder conditions. This lack of structural information hinders ligand discovery based solely on the receptor. A good alternative is a ligand-based strategy. Nevertheless, to date, much improvement is essential to allow more reliable *in silico* screening using this strategy.

This thesis is, therefore, timely and of great relevance as it proposes ligand-based prediction models for the development of novel computational tools and approaches capable of effectively predicting GPCRs ligands and supporting ranking of compounds on drug discovery investigations, which can be implemented and made freely available as user-friendly web servers. The development of new *in silico* GPCR ligand discovery methods will also provide important knowledge related to the link of chemical and biological aspects of this family of protein and its ligands, essential for the development of novel therapeutic options.

2 Aims

2.1 General Aim

The overall goal of this thesis was to develop a web-based platform (pdCSM-GPCR) that will provide support for the discovery of new active compounds for a comprehensive number of GPCRs using a ligand-based lead discovery approach. Besides, this study aimed to gather information about GPCRs ligand discovery and to gain a better understanding about the use of machine learning in the context of predicting bioactivity of small molecules.

2.2 Specific Aims

- Build up a database containing resourceful and curated of experimentally determined binding affinities for 36 different GPCRs, covering classes A, B1, C and F;
- Investigate molecular properties that compose known GPCR ligands to identify what makes up GPCR ligands;
- Develop predictive models capable of predicting bioactivity of small molecules when interacting with GPCRs;
- Design a user-friendly web-server containing all information derived from this thesis to support the discovery of novel leads through ranking compounds on drug discovery investigations, which would enable enriching screening libraries with compounds more likely to be active.

3 Methods

In this chapter, we describe the development of bioactivity predictive models for small molecules, aiming to target GPCRs (pdCSM-GPCR). We used the concept of applying graph-based structural signatures combined to auxiliary features as evidence for training machine learning algorithms. The combined features are correlated to the bioactivity values of actual molecule ligands and non-ligands (information obtained from PubChem) which produces predictors capable of inferring reliable bioactivity values on diverse molecules which were not used for training. The general pdCSM-GPCR workflow is depicted in Figure 6. It is composed of three main steps, including: (i) data set acquisition; (ii) feature engineering, and (iii) machine learning.

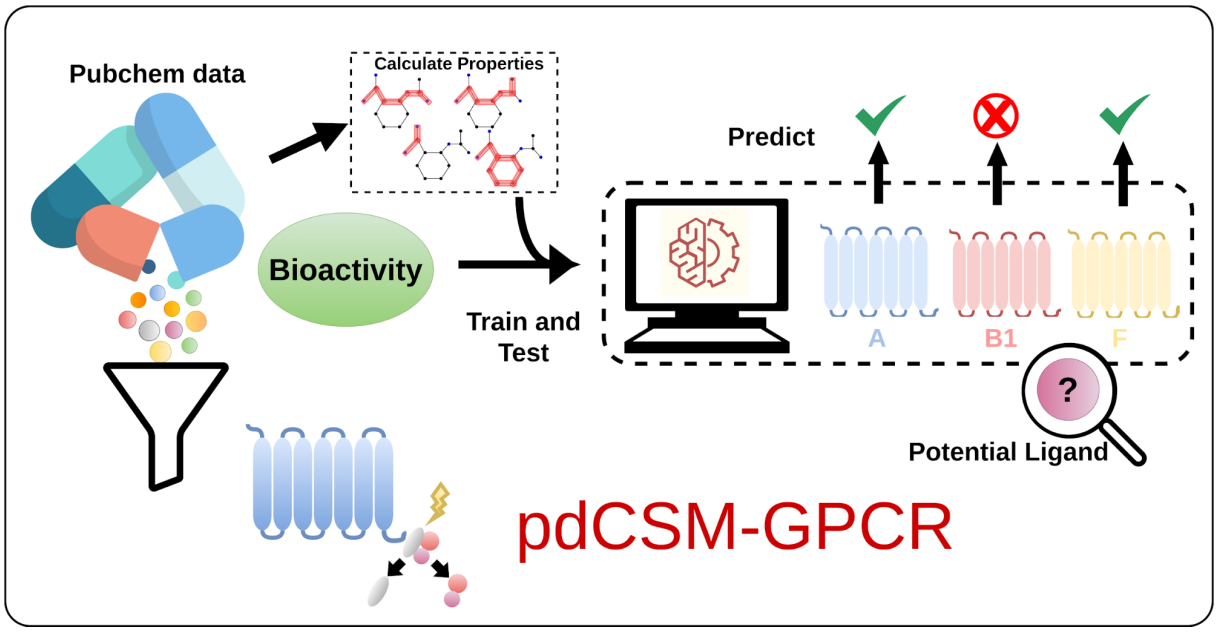


Figure 6 – pdCSM-GPCR workflow. Initially, we collected ligand data for 36 different GPCRs from PubChem (KIM et al., 2018), then we derived from them two types of features: compound auxiliary features (including molecular properties, toxicophores, and pharmacophores) and distance-based graph signatures. Afterwards, we used this information as the basis for the development of machine learning models for predicting bioactivity for GPCRs. This predictive models compose the web-based platform (pdCSM-GPCR) that enable ranking compounds on GPCRS drug discovery studies.

3.1 Data set acquisition

The collections were made by searching on the PubChem server, for the UniProt IDs belonging to GPCRs receptors of medical interest, according to the literature (see Table 4). We retrieved molecules binding affinities for 36 different GPCR receptors from PubChem, available in the section “Tested compounds” of the target (GPCR receptor) webpage. The PubChem CID

(Compound ID number) information available in the datasets was used to retrieve the SMILES (Simplified Molecular Input Line Entry Specification), also from PubChem. We choose to use CID, because different substance records (called SID in the PubChem) may contain different kinds of information for the same molecule, in order to place together this information there is a process called 'standardisation' that aggregates and stores them in the Compound database under the same identifier, CID (KIM et al., 2018).

SMILES is a chemical notation that allows representation of a chemical structure (WEININGER, 1988). They can represent using simple vocabulary (atom and bond symbols), and few grammar rules. 2-Propanol would be "CC(O)C" and 2-Methylbutanal would be "CC(C)CC(=O)". Using a type of SMILES called ISOMERIC SMILES it is even possible to represent specific isotopism, configuration about double bonds, and chirality. We applied the isomeric version to this thesis. Also regarding the SMILES, it is possible to represent the same molecule using different SMILES. We treated this cases using RDKit modules to generate the same SMILES string for a given molecule and avoid redundancy.

Subsequently, we filtered these datasets to be composed only by two columns, one comprising molecules represented as SMILES and a second column with their respective experimental bioactivity measurement in μM, which were converted to logarithmic scale for training ($-\log$[Molar])(see script A.1, Appendices). In pharmacology, bioactivity of compounds refers to a measure of potency (inhibition or activation) of a drug when interacting with a biological target. We only considered Ki, Kd, IC50, and EC50 for GPCRs as bioactivities as done in previous works (WU et al., 2018; WU et al., 2019; BURGGRAAFF et al., 2020; LIANG et al., 2019; KRUGER et al., 2014; ZIN; WILLIAMS; EKINS, 2020). We also filtered out repeated and dubious molecules (ligands with active and inactive status) for each bioactivity measure. These measures are detailed below.

Ki means inhibition constant, while Kd indicates dissociation constant. Both terms are used in biochemistry and pharmacology to report which binding affinity a small molecule or enzyme has for a receptor, enzyme, or other biomolecule. Kd possesses a more general meaning, since it quantifies the equilibrium between any type of ligand being free in solution and bound to a site in a protein, whereas Ki, necessarily, comprehends the binding equilibrium of an inhibitor to a biomolecule. IC50, in turn, means how much of a particular inhibitory drug is needed to inhibit a given biological process or biological component by 50%. Because it does not directly measure a binding equilibrium, IC50 is less precise than the previously mentioned bioactivities. Finally, EC50 stands for effective concentration at 50%. Which refers to the concentration of any type of drug at which 50% of its maximum effect is achieved, being a more general metric than IC50. These measurements are made through analytic procedures, which relies on the binding of ligand molecules to receptors. A signalling detection is used to determine the presence and extent of the ligand-receptor complexes formed. Usually, this signalling is determined electrochemically or through fluorescence detection.

The GPCRs classes covered by this work include four families (A, B1, C, and F) and 9 subfamilies, also including two receptors described as orphans. A complete view of the datasets used in this work can be obtained in Tables 3 and 4. In the next tables, class A receptors are coloured in blue, class B in green, class C in red and class F in purple.

3.2 Substructure mining

We evaluated which substructures of potent GPCR ligands were more frequent and absent in non-ligands molecules. For this task, the top 300 most potent ligands per receptor were selected or those with bioactivity greater than 5 (meaning potency of 10μM or higher potency - for receptor Q96LB2, only 87 molecules were selected). Molecular Substructure Miner (MoSS)(BORGELT; MEINL; BERTHOLD, 2005) was used to identify molecular substructures that were enriched in the group of potent ligands in comparison with the remainder of the data set.

MoSS finds molecular fragments that are frequent in a target part of the database, but rare in the complement part. For the GPCR ligand discovery task, we specified a support of 10% as minimum frequency for the group of potent ligands and a support of 2% as maximum frequency non-potent ligands of the data set. This means we are looking for fragments that appear with at least 10% in the group of potent ligands (and do not have super-structures that occur with the same frequency), but with no more than 2% as maximum frequency in the non-potent counterpart of the data set. The SMILES strings were input into the MoSS algorithm, for both potent GPCR ligands and the remainder molecules. Afterwards, we evaluated just molecular substructures exclusive for the potent ligands.

3.3 Feature engineering

Two main sets of molecular descriptors have been calculated based on the SMILES representation of the molecules and used in combination as evidence to train, test and validate machine learning methods for predicting GPCR ligands: (i) a distance-based graph signature and (ii) general features (general molecule chemical and topological property descriptors).

3.3.1 Graph-based and auxiliary signatures

Graph-based signatures compose a general representation of biological entities, their topology and chemical composition. They are a valuable representation for modelling biological entities, such as small molecules. In this thesis we used CSM, which are signatures modelled as unweighted, undirected graphs, where nodes represent atoms and edges represent covalent bonds.

Table 3 – Description of GPCRs considered in this work, with their respective families and subfamilies (class A receptors are coloured in blue, class B in green, class C in red and class F in purple).

Protein name	UniProt ID	Family	Subfamily
Muscarinic acetylcholine receptor M4	P08173	A	Aminergic
5-hydroxytryptamine receptor 1A	P08908	A	Aminergic
Muscarinic acetylcholine receptor M5	P08912	A	Aminergic
Muscarinic acetylcholine receptor M5	P0DMS8	A	Aminergic
Muscarinic acetylcholine receptor M3	P20309	A	Aminergic
Substance-K receptor	P21452	A	Peptide
D(4) dopamine receptor	P21917	A	Aminergic
Endothelin receptor type B	P24530	A	Peptide
5-hydroxytryptamine receptor 2C	P28335	A	Aminergic
Adenosine receptor A2b	P29275	A	Nucleotide
Adenosine receptor A1	P30542	A	Nucleotide
Gonadotropin-releasing hormone (type 1) receptor 1	P30968	A	Peptide
Prostaglandin E2 receptor EP1 subtype	P34995	A	Lipid
Somatostatin receptor type 5	P35346	A	Peptide
Alpha-1A adrenergic receptor	P35348	A	Aminergic
Mu-type opioid receptor	P35372	A	Peptide
B1 bradykinin receptor	P46663	A	Peptide
P2 purinoceptor subtype Y1	P47900	A	Nucleotide
Melatonin receptor type 1A	P48039	A	Peptide
5-Hydroxytryptamine receptor 6	P50406	A	Aminergic
C-C chemokine receptor type 3	P51677	A	Protein
Hydroxycarboxylic acid receptor 2	Q8TDS4	A	Alicarboxylic acid
G protein-coupled bile acid receptor 1	Q8TDU6	A	Steroid
Mas-related G protein-coupled receptor X1	Q96LB2	A	Orphan
Sphingosine 1-phosphate receptor 3	Q99500	A	Lipid
Melanin-concentrating hormone 1	Q99705	A	Peptide
Sphingosine 1-phosphate receptor 5	Q9H228	A	Lipid
G protein-coupled receptor 35	Q9HC97	A	Orphan
Histamine H3 receptor	Q9Y5N1	A	Aminergic
Prostaglandin D2 receptor 2	Q9Y5Y4	A	Lipid
Glucagon receptor	P47871	B1	Peptide
Calcitonin gene-related peptide type 1 receptor	Q16602	B1	Peptide
Extracellular calcium-sensing receptor	P41180	C	Ion
Metabotropic glutamate receptor 2	Q14416	C	Amino acid
Metabotropic glutamate receptor 4	Q14833	C	Amino acid
Smoothened homolog	Q99835	F	Protein

Table 4 – Description of GPCRs considered in this work: Medical importance and number of compounds with available bioactivity (class A receptors are coloured in blue, class B in green, class C in red and class F in purple).

Protein name	Medical interest	#Ligands collected
Muscarinic acetylcholine receptor M4	Parkinson's disease (https://doi.org/10.1016/j.pharmthera.2007.09.009).	1097720
5-hydroxytryptamine receptor 1A	Neuropsychiatric disorders such as anxiety, depression, and schizophrenia.	135544
Muscarinic acetylcholine receptor M5	Tobacco and cannabis dependence (10.1186/1471-2156-8-46).	1097830
Muscarinic acetylcholine receptor M5	Rheumatoid arthritis.	10929
Muscarinic acetylcholine receptor M3	Type 2 diabetes (https://doi.org/10.1016/j.cmet.2006.04.009)	6786
Substance-K receptor	Inflammatory and pain responses (https://doi.org/10.1016/j.neulet.2005.06.011).	3153
D(4) dopamine receptor	Parkinson's disease, schizophrenia, mania, depression, substance abuse, and eating disorders (https://doi.org/10.1021/cr050263h).	6251
Endothelin receptor type B	Hirschsprung's disease (10.1074/jbc.273.18.11378).	1805
5-hydroxytryptamine receptor 2C	Neuroendocrine responses to stress (10.1523/JNEUROSCI.2584-06.2007)	8179
Adenosine receptor A2b	Asthma and gastrointestinal disorders (https://doi.org/10.1016/B978-0-12-803724-9.00001-6).	6714
Adenosine receptor A1	Cardiac ischemia, stroke, hypertension, and epilepsy.	13364
Gonadotropin-releasing hormone (type 1) receptor 1	Hypogonadotropic hypogonadism (https://doi.org/10.1038/ng0198-14).	3017
Prostaglandin E2 receptor EP1 subtype	Treatment of neuropathic pain (10.1097/00000539-200110000-00043).	1631
Somatostatin receptor type 5	Inhibit the release of many hormones and other secretory proteins (10.1159/000054651).	1361
Alpha-1A adrenergic receptor	Noradrenergic modulation of olfactory driven behaviours (10.1113/jphysiol.2012.248591).	4034
Mu-type opioid receptor	Morphine-induced analgesia and itch (10.1016/j.cell.2011.08.043).	691466
B1 bradykinin receptor	Inflammatory injuries that follow ischaemia and reperfusion (10.4049/jimmunol.172.4.2542).	1491
P2 purinoceptor subtype Y1	Platelet shape and platelet aggregation (10.1042/bj3360513).	1200
Melatonin receptor type 1A	Circadian and neuroendocrine disorders (0.1006/geno.1995.1056).	3003
5-Hydroxytryptamine receptor 6	Learning process and memory (https://doi.org/10.1016/B978-0-12-800836-2.00011-8).	8230
C-C chemokine receptor type 3	Binds and responds to a variety of chemokines, HIV infection (10.1016/s0092-8674(00)81313-6 0).	1675
Hydroxycarboxylic acid receptor 2	Dyslipidemia (10.2217/pgs.15.79).	1664
G protein-coupled bile acid receptor 1	Immune and inflammatory liver diseases (10.1002/hep.24525).	1014
Mas-related G protein-coupled receptor X1	Modulation of nociception (https://doi.org/10.1096/fj.202001667RR).	936090
Sphingosine 1-phosphate receptor 3	Glioblastoma (10.1016/j.freeradbiomed.2005.09.015).	232193
Melanin-concentrating hormone receptors 1	Obesity (10.2174/092986708784049621).	8628
Sphingosine 1-phosphate receptor 5	Huntington's disease (10.1093/hmg/ddy153).	782
G protein-coupled receptor 35	Albright hereditary osteodystrophy-like phenotype (10.1111/j.1399-0004.2004.00363.x).	293497
Histamine H3 receptor	Attention deficit hyperactivity disorder, Alzheimer's disease and schizophrenia (10.1038/bjp.2008.147).	6873
Prostaglandin D2 receptor 2	Inflammatory disease of the upper airways (10.1016/j.prostaglandins.2003.12.002).	5017

Protein name	Medical interest	#Ligands collected
Glucagon receptor	Type 2 diabetes (10.1038/ng0395-299).	2053
Calcitonin gene-related peptide type 1 receptor	Migraine (10.1177/1756285610388343).	1663
Extracellular calcium-sensing receptor	Ischemic brain injury (10.1002/acn3.118).	718
Metabotropic glutamate receptor 2	Pain mechanisms and behavioral modulation (https://doi.org/10.3389/fnmol.2018.00383).	2475
Metabotropic glutamate receptor 4	Parkinson Disease (10.1007/s11481-016-9655-z).	3457
Smoothened homolog	Carcinogenesis (https://doi.org/10.1016/j.lfs.2020.117302).	1366

In order to calculate them, firstly, the shortest distance between all atom pairs (nodes) are calculated. These distances are measurements of the shortest paths between two atoms, according to the covalent bonds between them, one of distance meaning one covalent bond. Following this step, according to a defined range of distances (called cutoffs and defined by the sum of the bonds between the pair of atoms) and a distance step, the molecule is scanned through these distances, computing the frequency of type of atom pairs (categorised by pharmacophore type), that are close according to this distance threshold. CSM generates feature vectors that represent distance patterns between atoms. The motivation for using those signatures lies in the fact that diferent molecules will generate different distributions of distances between their atoms and, consequently, pharmacophoric groups. This information is captured by the CSM and can be used as evidence for ML models. This mentioned step was already successfully employed at pkCSM (predicting small-molecule pharmacokinetic properties- (PIRES; BLUNDELL; ASCHER, 2015), DUET (predicting effects of mutations on protein stability- (PIRES; ASCHER; BLUNDELL, 2014), mCSM (predicting the effect of mutations in proteins- (PIRES; ASCHER, 2017), aCSM (receptor-based ligand prediction- (PIRES et al., 2013), CSM-Lig (assessing and comparing protein–small molecule affinities- (PIRES; ASCHER, 2016a).

In this work, we vary the distance threshold from 5 bonds to 20 bonds, with “one bond” as a distance step. We scanned through these distance thresholds, computing the frequency of atom pairs that are close according to the mentioned distance threshold. Together, these vectors compose the CSM signature. Each line of the matrix represents one molecule, and each column represents the frequency of pharmacophoric atom pairs within a certain distance (see Figure 7) for details about distance-based graph signatures).

3.3.2 Auxiliary features

Auxiliary signatures refer to various molecular properties describing the general physicochemical properties of compounds (TODESCHINI; CONSONNI, 2009). These signatures include toxicophore fingerprints proposed by (KAZIUS; MCGUIRE; BURSI, 2004), atomic pharmacophore frequency count, and general molecular properties including lipophilicity (log P), molecular weight, surface area, number of rotatable bonds. The toxicophore fingerprint was calculated based on substructure matching from SMILES arbitrary target specification (SMARTS) queries published by (KAZIUS; MCGUIRE; BURSI, 2004), and the other auxiliary signatures

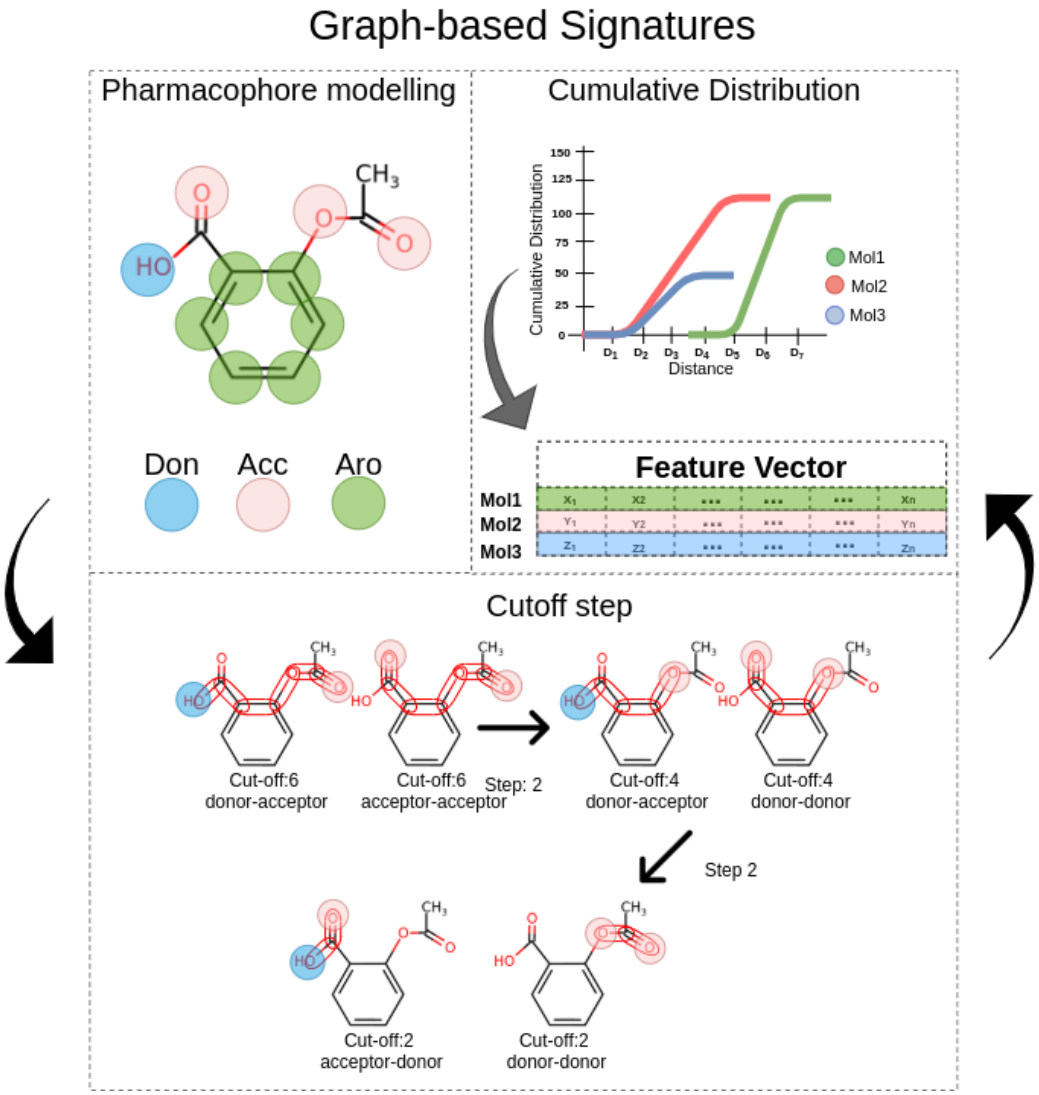


Figure 7 – Modelling small molecule activity using graph-based signatures. We picture on the image a small molecule being scanned for two types of pharmacophores, hydrogen bond acceptor and donor. We started with a distance cut off of 6 bonds, and found two pairs of pharmacophores, donor-acceptor and acceptor-acceptor. Following a distance step of 2, we then used a distance cutoff of 4 and found also two pharmacophores, donor-acceptor, donor-donor. At last, we used a cut-off of 2 and found also two pairs of pharmacophores, acceptor-donor, donor-donor (bottom panel). The small molecule is then represented as cumulative distributions of the pair of pharmacophores (top-right panel).

were calculated using the RDKit cheminformatics toolkit[1] . A complete list of auxiliary features used in this thesis is described in Table 5.

3.4 Machine Learning Algorithms

Prediction of compound bioactivities was framed as a regression task for predicting bioactivities (Ki, Kd, IC50, and EC50), with a range of different supervised learning algorithms being assessed, including Extra Trees (GEURTS; ERNST; WEHENKEL, 2006), Random Forest (BREIMAN, 2001), Gradient Boost (FRIEDMAN, 2001) and XGBoost (CHEN; GUESTRIN, 2016) regressors. All these four algorithms belong to the class of ensembles of decision trees. These methods combine multiple ML models to create more powerful models. And, all of them use decision trees as building blocks. In essence, decision trees are composed of if/else questions disposed in a hierarchical manner, following these questions the model is capable of reaching a decision (GUIDO; MUELLER, 2016). In the case of our question, the actual output is a numeric value of the bioactivity. The decision to reach prediction is based on the features (graph based signatures and auxiliary features) we used as input for the ML algorithm.

Random Forest implements many decision trees, each one is different from the others, the trees are built at random, considering different features (and also samples). Each tree is capable of doing proper predictions, but tends to overfit on part of the data. When many trees are combined, overfitting will be reduced by averaging the tree results. The randomness from this algorithm comes from the fact that the trees are built in a randomised way: by the selection of different data points and different features in each split test (GUIDO; MUELLER, 2016).

The Gradient Boost, in contrast to the Random Forest approach, creates serialised trees, where each tree tries to correct the mistakes of the previous one. The trees created on this algorithm are shallow and present low depth so that the models are capable of providing good predictions on part of the data. The combination of these simple models can generate models capable of predicting with great reliability (FRIEDMAN, 2001; GUIDO; MUELLER, 2016).

Extra trees are very similar to Random Forest, the main difference lies in the fact that instead of computing the locally optimal feature/split combination, for each feature under consideration, this algorithm selects its cut-point fully at random, independently of the target variable (GEURTS; ERNST; WEHENKEL, 2006).

XGBoost stands for eXtreme Gradient Boosting. It follows the principle of gradient boosting, however, uses a more regularised model formalisation to control over-fitting (CHEN; GUESTRIN, 2016).

The best performing models were selected based on Pearson's, Spearman's and Kendall's correlation coefficients and Root Mean Square Error (RMSE). The Scikit-learn library (version

[1] "RDKit." https://www.rdkit.org/.

Table 5 – Auxiliary features.

Name	Description	Reference
HeavyAtomCount	Number of Non-Hydrogen atoms in a given molecule	
LogP	Particular ratio of the solute concentrations between the two solvents (a biphase of liquid phases), one of the solvents is water and the other is a non-polar solvent	(Wildman and Crippen, 1999)
NumHeteroatoms	Number of heavy atoms a molecule. (Non-hydrogens)	
NumRotatableBonds	Number of Rotatable Bonds	
RingCount	Number of rings.	
TPSA	Topological polar surface area (TPSA) of a molecule is defined as the surface sum over all polar atoms, primarily oxygen and nitrogen, also including their attached hydrogen atom.	(Ertl et al., 2000)
LabuteASA	Labute's Approximate Surface Area	(Labute, 2000)
MolWt	Molecular Weight	
Fcount		
Tox	Toxicophores	(Kazius et al., 2005)
BalabanJ:	Balaban's connectivity topological index	(Balaban, 1982)
BertzCT	A topological index meant to quantify "complexity" of molecules. Consists of a sum of two terms, one representing the complexity of the bonding, the other representing the complexity of the distribution of heteroatoms.	(Bertz, 1981)
Chi0, Chi1	Atomic connectivity index (order 0) . This is calculated as the sum of 1/sqrt(di) overall heavy atoms i with di >0.	(Hall and Kier, 2007)
Chi0n - Chi4n		(Hall and Kier, 2007)
Chi0v - Chi4v	Atomic connectivity index (order 1). This is calculated as the sum of 1/sqrt(didj) overall bonds between heavy atoms i and j where i <j.	(Hall and Kier, 2007)
chi0v_C, chi1v_C	Carbon valence connectivity index (order 0). This is calculated as the sum of 1/sqrt(vi) overall carbon atoms i with vi >0.	(Hall and Kier, 2007)
HallKierAlpha		(Hall and Kier, 2007)
Kappa1- Kappa3		(Hall and Kier, 2007)
PEOE_VSA1 - PEOE_VSA14	MOE-type descriptors using partial charges and area contributions	
SMR_VSA1 - SMR_VSA10	MOE-type descriptors using MR contributions and surface area contributions	
SlogP_VSA1 - SlogP_VSA12	MOE-type descriptors using SLogP contributions and surface area contributions	
EState_VSA1 - EState_VSA11	MOE-type descriptors using EState indices and surface area contributions	
VSA_EState1 - VSA_EState10	MOE-type descriptors using surface area contributions and Estate indices	
Organic functions	Al_COO, Al_OH, Al_OH_noTert, ArN, Ar_COO, Ar_N, Ar_NH, Ar_OH, COO, COO2, C_O, C_O_noCOO, C_S, HOCCN, Imine, NH0, NH1, NH2, N_O, Ndealkylation1, Ndealkylation2, Nhpyrrole, SH, aldehyde, alkyl_carbamate, alkyl_halide, allylic_oxid, amide, amidine, aniline, aryl_methyl, azide, azo, barbitur, benzene, benzodiazepine, bicyclic,f_diazo, dihydropyridine, epoxide, ester, ether, furan, guanido, halogen, hdrzine, hdrzone, imidazole, imide, isocyan, isothiocyan, ketone, ketone_Topliss, lactam, lactone, methoxy, morpholine, nitrile, nitro, nitro_arom, nitro_arom_nonortho, nitroso, oxazole, oxime, para_hydroxylation, phenol, phenol_noOrthoHbond, phos_acid, phos_ester, piperdine, piperzine, priamide, prisulfonamd, pyridine, quatN, sulfide, sulfonamd, sulfone, term_acetylene, tetrazole, thiazole, thiocyan, thiophene, unbrch_alkane, urea	

0.20.3) for Python (version 2.7) (PEDREGOSA et al., 2011) was used for training and testing the models. For all ML algorithms, the parameter "random_state" was set to 0. This parameter controls the random seed given to each Tree estimator at each boosting iteration. Random seed is a number (or a vector) used to initialise a pseudorandom number generator. When keeping the same one, we guaranteed that we get the same training and validation data set through all machine learning experiments. All other parameters were kept as default and no hyperparameter tuning was performed.

3.5 Performance metrics

In order to check the statistical relevance of our results, we used the Pearson correlation. Pearson correlation is a measure of the linear correlation between two variables X and Y. It ranges from -1 to 1. Correlations of -1 or +1 indicate a perfect linear correlation. Positive correlations point that as X increases, so does Y. Negative correlations point that as X increases, Y decreases. A value of 0 implies that there is no linear correlation between X and Y (experimental molecular bioactivity vs. predicted bioactivity). In our case, one variable is the molecular bioactivity obtained experimentally, and the other variable is the predicted bioactivity. The larger the r-value is, the better the model performance will be. The formula for the Pearson correlation is:

$$r = \frac{\sum(x - m_x)(y - m_y)}{\sqrt{\sum(x - m_x)^2 \sum(y - m_y)^2}}$$

Where mx is the mean of the x (predicted bioactivity) vector and my is the mean of the vector y (experimental bioactivity).

Additionally, we used Spearman's correlation coefficient, Kendall's correlation coefficient, and Root Mean Square Error. Spearman's correlation coefficient is a nonparametric measure of rank correlation. It assesses how well the relationship between two variables can be described using a monotonic (whether linear or not) function. The formula for the Spearman's correlation is:

$$r_s = \rho_{\mathrm{R}(X),\mathrm{R}(Y)} = \frac{\mathrm{cov}(\mathrm{R}(X), \mathrm{R}(Y))}{\sigma_{\mathrm{R}(X)} \sigma_{\mathrm{R}(Y)}},$$

ρ denotes the usual Pearson correlation coefficient, but applied to the rank variables. $\mathrm{cov}(\mathrm{R}(X), \mathrm{R}(Y))$, is the covariance of the rank variables, and $\sigma_{\mathrm{R}(X)}$, and $\sigma_{\mathrm{R}(Y)}$ are the standard deviations of the rank variables.

Kendall's correlation coefficient is also a nonparametric measure of the strength and direction of association that exists between two variables. The Kendall coefficient is defined as:

$$\tau = \frac{(\text{number of concordant pairs}) - (\text{number of discordant pairs})}{\binom{n}{2}}.$$

Where $\binom{n}{2} = \frac{n(n-1)}{2}$ is the binomial coefficient for the number of ways to choose two items from n items.

RMSE is a measure of the differences between actual and predicted values. It represents a square root of the differences between predicted values and observed values:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (Predicted_i - Actual_i)^2}{N}}$$

where, Predicted$i$, means the predicted values, Actual$i$, means the observed (actual) values, and N, means total number of observations.

3.6 Model validation

In order to test and validate our models internally, we used cross-validation. Through the model validation step, it is possible to estimate how accurately a predictive model will perform in practice, and avoid overfitting, which is when a model function is too closely fit the training dataset only and is not reliable in predicting unseen data during the training. In this approach, called k-fold cross-validation, the training set is split into k smaller sets. Then for each of the k “folds”: the model is trained using k - 1 folds as training data and the resulting model is validated on the remaining part of the data. The performance is defined as the average of the results computed in the loop. For each model, we employed stratified 5-, 10- and 20-fold cross-validation on the training set. It enabled us to evaluate how the results of the ML models were capable of being generalised to an independent dataset. Performance was also assessed on 90% of the data, after removing 10% of the worst predicted data point, to evaluate the effects of outliers in model prediction capabilities. This removal of 10% of outliers was done just for the sake of performance analysis, in the end we used all molecules for generating the final models. The Scikit-Learn toolkit was used in all ML procedures.

We also tested our models using low-redundancy independent blind test sets. For this purpose, datasets per GPCR were split into training (90%) and blind tests (10%). The split between test and training data was done using a Python algorithm (see script A.2, Appendices) that first clusters by similarity all the data according to Morgan fingerprints. In the clustering step, we used the Butina clustering algorithm (BUTINA, 1999) from RDKit and a cutoff of 80% of similarity, meaning that the molecules inside a cluster had at least 80% of similarity.

For similarity comparison, we used Tanimoto coefficient, which is defined as the ratio of the intersection of the two sets over the union of the two sets (sets in this case, meaning fingerprints generated for each SMILE). After the generation of clusters, they were randomly selected, and grouped to form the test group with approximately 10% of the molecules, the other part was used as a training set.

3.7 Feature selection

As a means of finding the most relevant features for the regression of GPCR ligand problems and removing the irrelevant features, we performed a feature selection using all features (distance-based graph signatures and auxiliary features). Feature selection is an automatic procedure of reducing the number of features when developing a machine learning predictive model. It is desirable because it both reduces the computational cost of modelling and, also enables in many cases, the improvement of the predictive performance of the model, avoiding overfitting. For this task, we used a greedy feature selection algorithm (see Figure 8) which is a heuristic algorithm that aims to reach a global optimum solution by making locally optimal choices at each stage. The adopted algorithm employs a forward selection. At its first step, greedy feature selection tries all features individually, fixing the one with the best score. At the second step, all features that are left are tested with the fixed one and then, the second-best feature is fixed. Subsequently, these steps happen as long as the performance improves (Pearson correlation meliorates at any value, not cutoff was applied in this step). Concerning the feature selection, we used 10-fold cross validation for selecting the best case, choosing the higher Pearson correlation coefficient.

3.8 Performance comparison with alternative methods

We compared our predictive model's performances with WDL-RF (WU et al., 2018). The comparison was done using the data sets provided by the authors while training their models available online. Initially, SMILES for each data set were submitted to the WDL-RF web server. The web server outputs a table containing a column with SMILES and another with the predicted bioactivity in nanoMolar. Predictions were converted to a standard value using $-\log10$(bioactivity in nanoMolar), consistent with what was performed by (WU et al., 2018). The same procedure was employed using our web server.

WDL-RF predictors were built using weighted deep learning and random forest, to model the bioactivity. They consisted of two consecutive rounds: (i) generation of fingerprints using a weighted deep learning method, followed by (ii) bioactivity calculations using random forest models. These predictors comprehend 26 GPCRs.

We also checked the performance of our models, using "control ligands", as reported by

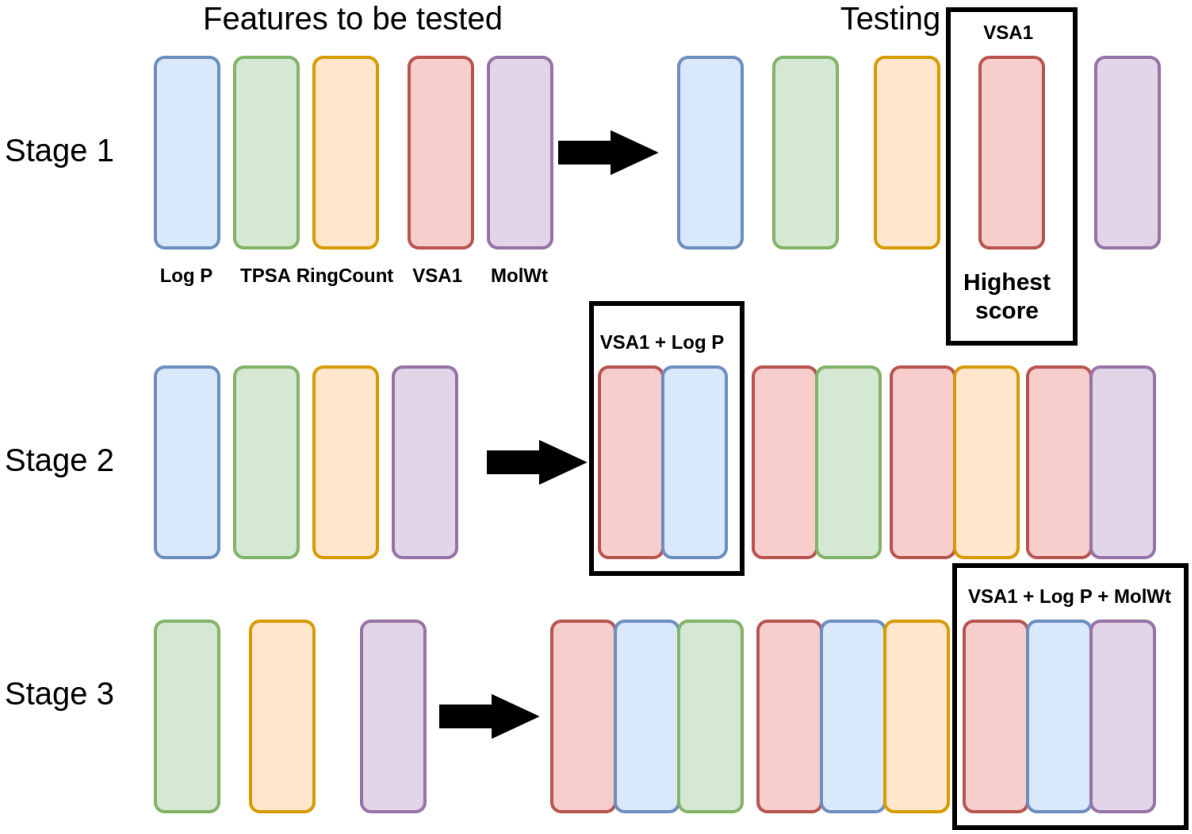


Figure 8 – Feature Selection. All features are tried individually, then the best one is fixed. At the second stage test, all features that are left are tested again combined with the fixed one and afterwards, the second-best feature is fixed. Subsequently, it continues iteratively to happen as long as the performance improves.

(WU et al., 2018). For this, we included a 'non-ligand' set, in order to check if the presence of these control ligands would increase performance of our models. The small molecules for these 'non-ligand' sets were obtained through DUD-E (MYSINGER et al., 2012), a tool that generates decoys (non-ligands molecules) using active compounds. For this purpose, we used top potent ligands from our datasets, the same we selected in the section "Substructure mining". We added to our datasets 20% of decoys and the bioactivity of these were set to -1 (10 molar).

3.9 Website Design and Implementation

The pdCSM-GPCR web server was designed to provide a user-friendly, reliable and scalable web interface to predict bioactivity for GPCR ligands (<http://biosig.unimelb.edu.au/pdcsm_gpcr/>). It was implemented using Bootstrap 3.3.7 and Flask framework, version 1.1.2. The 2D chemical structure depictions on the web server are generated with RDKit.

4 Results

In this chapter, we present new bioactivity predictors for the study of 36 different GPCRs belonging to four classes (A, B1, C, and F). We devised a range of experiments in order to better understand and contrast the molecular properties of ligands targeting different GPCRs, demonstrate the accuracy of pdCSM-GPCR models and compare their performance with other available methods.

4.1 Data sets

Initially we retrieved small-molecule bioactivities for 26 different GPCRs, covering four major classes, from PubChem (KIM et al., 2018). These were done for the sake of performing a direct comparison with a previous method, WDL-RF (WU et al., 2018). We, however, have further expanded this set by curating more data from the literature to include seven new datasets for class A, a new predictor for the B1 class (UniProt ID: Q16602), one for class C (UniProt ID: Q14833) and one for an orphan GPCR (UniProt ID:Q96LB2). In total, bioactivity data for 36 different GPCRs were collected, making this the most comprehensive dataset to date. Most part of the data set comprises class A receptors. This is due to the fact that the most targeted GPCR class has historically been class A and also this class accounts for nearly 80% of GPCR genes (Davies et al., 2007 (DAVIES et al., 2007).

We analysed and curated the retrieved data sets (see Table 6). These datasets included in most of the cases a range of different experimental studies. Because of that, some ligands presented duplicated results and also discrepancies. In case of two or more entries for the same molecule, only the first one to appear in the dataset was kept (neither bioactivity value nor class were considered for this step). In order to improve our training, we remove these inconsistencies using a Python script (see script A.1, Appendices).

Some receptors, such as Muscarinic acetylcholine receptor M4-(P08173), had a significant cutback in the amount of data after the curation step. These happened because great part of the data was actually qualitative, without the actual value of bioactivity, what is essential for regression models. Usually during the screening experiments, many compounds are screened using assays that are qualitative. Then the promising ones are tested quantitatively.

Figure 9 depict activity distributions on all datasets after filtering. It is important to point out that for most of the receptors, the bioactivity distribution range spans from four to more or less 15, where bioactivity is defined by -log10(activity in Molar). The majority of the molecules have a bioactivity between 6 and 8, according to the average and the median. (ZHANG et al., 2015), who performed a rigorous evaluation of ligands bound to elucidated structures of GPCRs,

Table 6 – Characteristics of the GPCRs datasets before and after filtering, and also the number of molecules in the group used for machine learning training and testing purposes (blind test validation) (class A receptors are coloured in blue, class B in green, class C in red and class F in purple. Lig collec= Number of collected ligands, Total aft. filt.= Total number of molecules after filtering, Train = Training set of ligands, Test= Test set of ligands).

Protein name	UniProt ID	#Lig collec.	#Total aft. filt.	#Train	#Test
Muscarinic acetylcholine receptor M4	P08173	1097720	978	837	141
5-hydroxytryptamine receptor 1A	P08908	135544	3790	3370	420
Muscarinic acetylcholine receptor M5	P08912	1097830	959	820	139
Muscarinic acetylcholine receptor M5	P0DMS8	10929	3513	3100	413
Muscarinic acetylcholine receptor M3	P20309	6786	2008	1698	310
Substance-K receptor	P21452	3153	922	762	160
D(4) dopamine receptor	P21917	6251	2335	2059	276
Endothelin receptor type B	P24530	1805	987	815	172
5-hydroxytryptamine receptor 2C	P28335	8179	3118	2765	353
Adenosine receptor A2b	P29275	6714	2109	1835	274
Adenosine receptor A1	P30542	13364	3833	3409	424
Gonadotropin-releasing hor. t1 receptor 1	P30968	3017	1373	1097	276
Prostaglandin E2 receptor EP1 subtype	P34995	1631	741	640	101
Somatostatin receptor type 5	P35346	1361	747	576	171
Alpha-1A adrenergic receptor	P35348	4034	1898	1645	253
Mu-type opioid receptor	P35372	691466	5275	4651	624
B1 bradykinin receptor	P46663	1491	756	608	148
P2 purinoceptor subtype Y1	P47900	1200	568	461	107
Melatonin receptor type 1A	P48039	3003	1043	891	152
5-Hydroxytryptamine receptor 6	P50406	8230	3044	2699	345
C-C chemokine receptor type 3	P51677	1675	1131	947	184
Hydroxycarboxylic acid receptor 2	Q8TDS4	1664	504	434	70
G protein-coupled bile acid receptor 1	Q8TDU6	1014	443	372	71
Mas-related G protein-coupled receptor X1	Q96LB2	936090	93	70	23
Sphingosine 1-phosphate receptor 3	Q99500	232193	1088	939	149
Melanin-concentrating hormone receptors 1	Q99705	8628	3721	3286	435
Sphingosine 1-phosphate receptor 5	Q9H228	782	417	349	68
G protein-coupled receptor 35	Q9HC97	293497	480	408	72
Histamine H3 receptor	Q9Y5N1	6873	3597	3133	464
Prostaglandin D2 receptor 2	Q9Y5Y4	5017	2749	2407	342
Glucagon receptor	P47871	2053	1006	843	163
Calcitonin gene-related peptide t1 receptor	Q16602	1663	757	612	145
Extracellular calcium-sensing receptor	P41180	718	535	439	96
Metabotropic glutamate receptor 2	Q14416	2475	1168	1025	143
Metabotropic glutamate receptor 4	Q14833	3457	579	504	75
Smoothened homolog	Q99835	1366	718	603	115

stated that ligand affinity in solved GPCR structures, generally is a single-digit nM range value. When we convert a single-digit nM (1e-9 Molar) values using -log10(activity in Molar), we got a value of 9 approximately. According to this assumption, a bioactivity (logarithmized value, higher means higher potency) between 6 or 8 would imply no or lower activity, meaning that datasets covered ligands and non-ligands.

One interesting case was the G protein-coupled receptor 35 (Q9HC97), which displayed a very different activity distribution. Most of the dataset molecules from this data set featured a bioactivity (-log10(activity in Molar)) equal four, which represents a very low affinity value. This receptor was first identified in 2000 and has been extensively studied for treatment of inflammatory bowel disease. However, it officially remains defined as an “orphan” GPCR (QUON et al., 2020). This hardship in discovering natural ligands could be linked to the fact that

this receptor would need different ligands to be activated, one orthosteric and other allosteric. Thus, high throughput essays, with one type of molecules per time, would indicate low affinity.

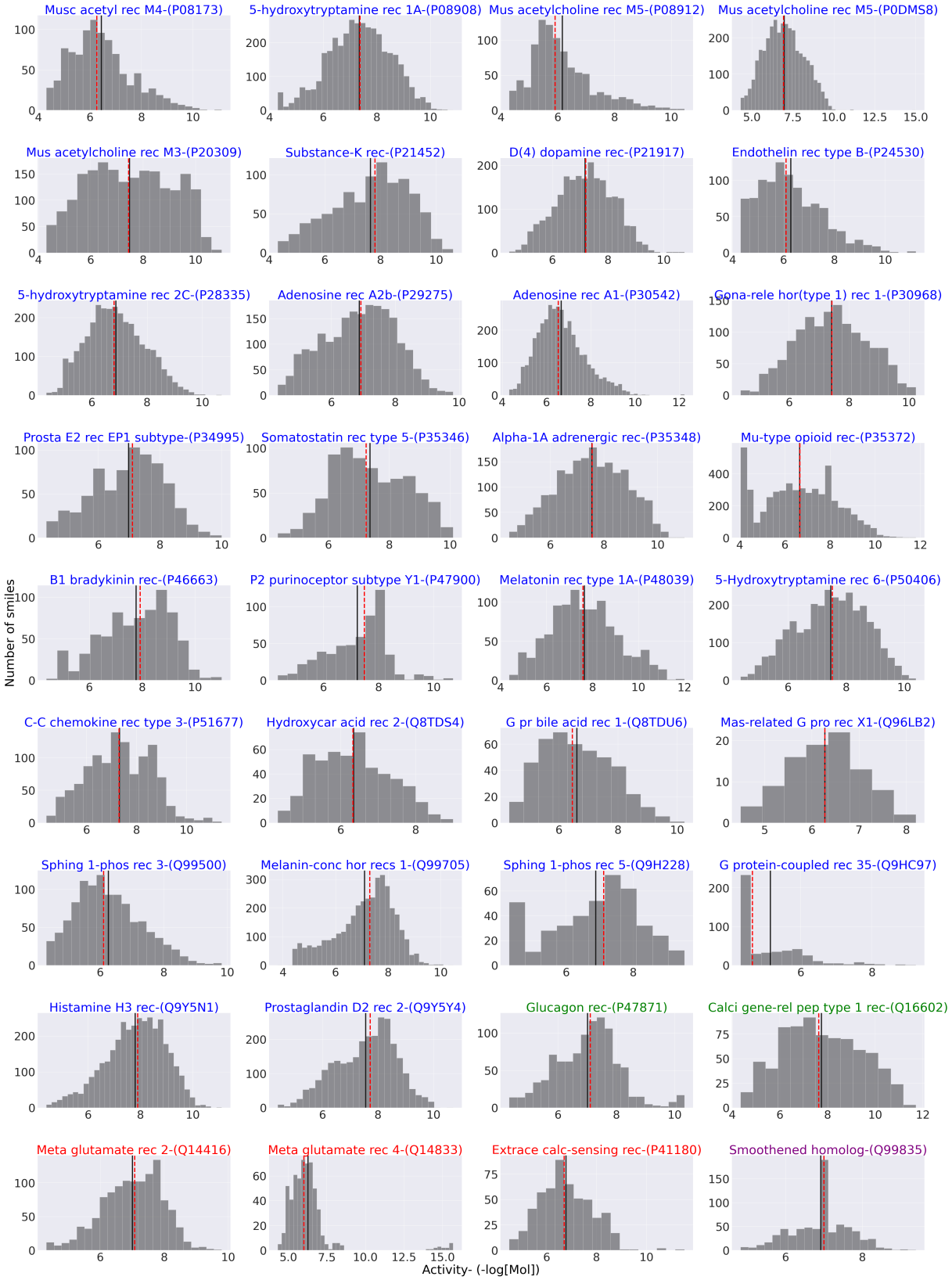


Figure 9 – Value of activity distributions - Datasets - At the x-axis the bioactivity is represented as -log[concentration], concentration in molar value. And, at the y-axis, the number of molecules is represented. The red line represents the median and the black line represents the average.

4.2 Analysis of molecular properties: what makes a GPCR ligand?

The top 300 most potent molecules (the average of 15% of all datasets) per receptor were selected or those with bioactivity greater than 5 (meaning potency of 10μM or higher). Using this data, we evaluated common molecular substructures of potent GPCR ligands. MoSS (BORGELT; MEINL; BERTHOLD, 2005) was used to identify molecular substructures that were enriched in the group of potent ligands in comparison with the remainder of the data set (Figure 10). Aromatic rings and nitrogen containing substructures were amongst the most enriched substructures in potent GPCR ligands across all classes. Similar results were found by (HORST et al., 2009). Applying substructure mining of GPCR ligands, they found that alkane amine substructures were enriched in the pool of substructures when comparing to non ligands, they also stated that these substructures are often linked to an aromatic system. Two other studies also contemplated the importance of amine substructures. One of these studies was (STRADER et al., 1988), they studied a beta-adrenergic receptor. They stated that a salt bridge between the ligands' protonated amino group and a negatively charged aspartic acid residue in transmembrane 3 is essential for the interaction happens. The other study, from (KOOISTRA et al., 2013), also stated that a negative charge in the residue $D^{3.32}$ in opioid receptors plays an important role in binding of positively ionised ligands via ionic interactions. These substructures could be correlated with key interactions between ligand and transmembrane parts of GPCRs and should be considered during ligand screening of large compound libraries and also for lead optimisation. They can be used to do a pre-screening of molecules in larger datasets or for evaluation of potential leads. On the other hand, it is also critical to note that, as these ligands characteristics were common for different types of GPCRs, they can bind more than one target. Thus, they should be carefully evaluated during drug discovery for GPCRs, for avoiding multiple GPCRs activation and drug collateral effects.

Besides looking for molecular substructures, we also found a limited number of potent molecules being shared among at least different GPCRs (see Figure 11) (no control regarding other family of receptors was made). The 21 ligands identified were shared between class A receptors and, in general, their properties (substructures) were consistent with what we observed for the most potent ligands across different receptors. They all possessed aromatic rings and substructures with nitrogens.

We also assessed common physicochemical properties of these potent ligands (Figures A1 to A8 of Appendices). We found that most potent ligands possessed between 20 and 40 heavy atoms, had a molecular weight between 200-500 daltons, less than 10 rotatable bonds, a polar surface area no greater than 140 $Å^2$ and a logP between 0-6 range (GHOSE; VISWANADHAN; WENDOLOSKI, 1998). The most potent compounds also possess between 2 and 12 heteroatoms, between 2 and 6 rings and a LabuteASA in the range of 150 to 200 $Å^2$. This is largely con-

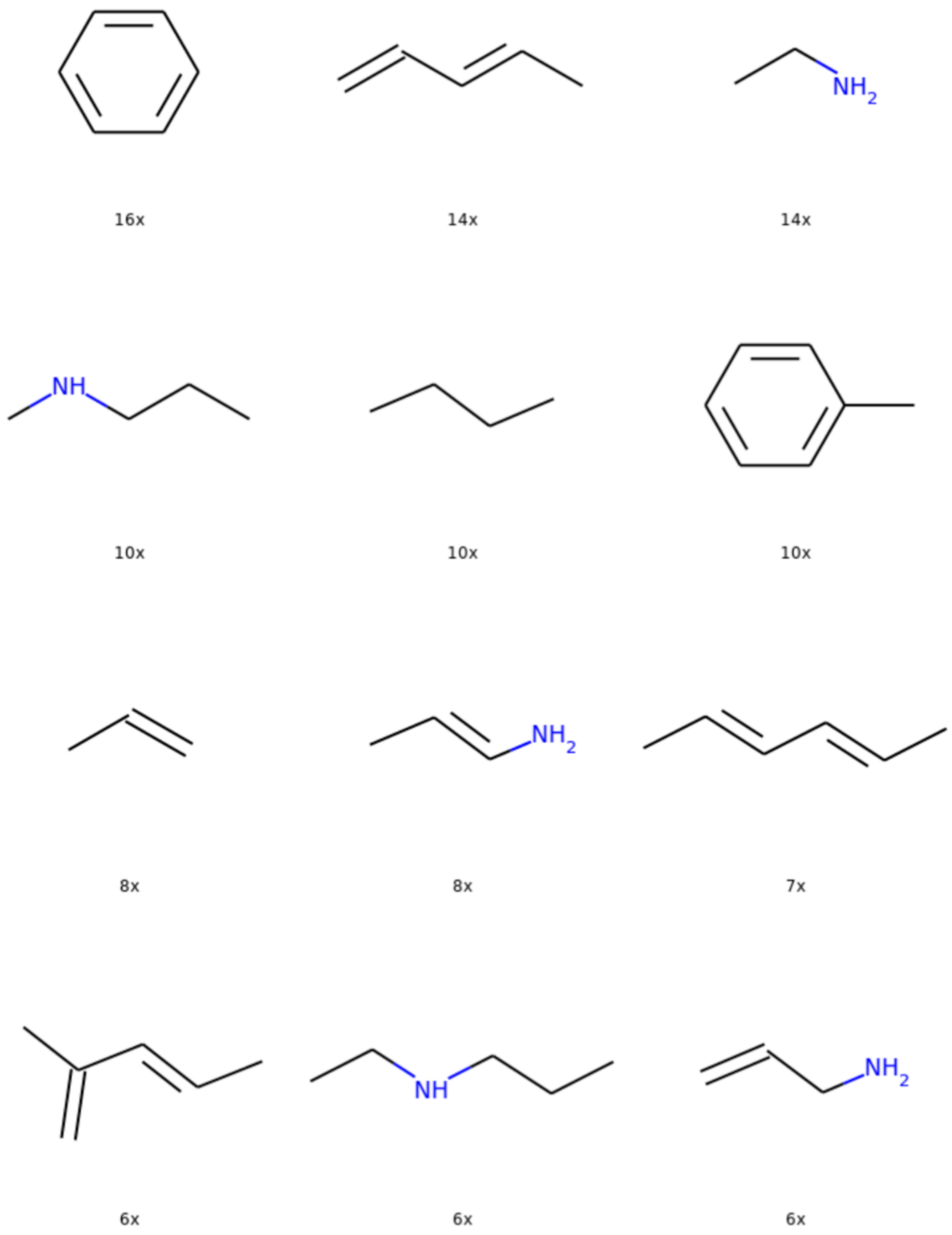


Figure 10 – Distribution of the top ten most frequent substructures present on the most active ligands- This distribution comprise the data sets of all receptors. The number below the fragment refers to how many receptors (data sets) had that fragment enriched in the most potent molecules.

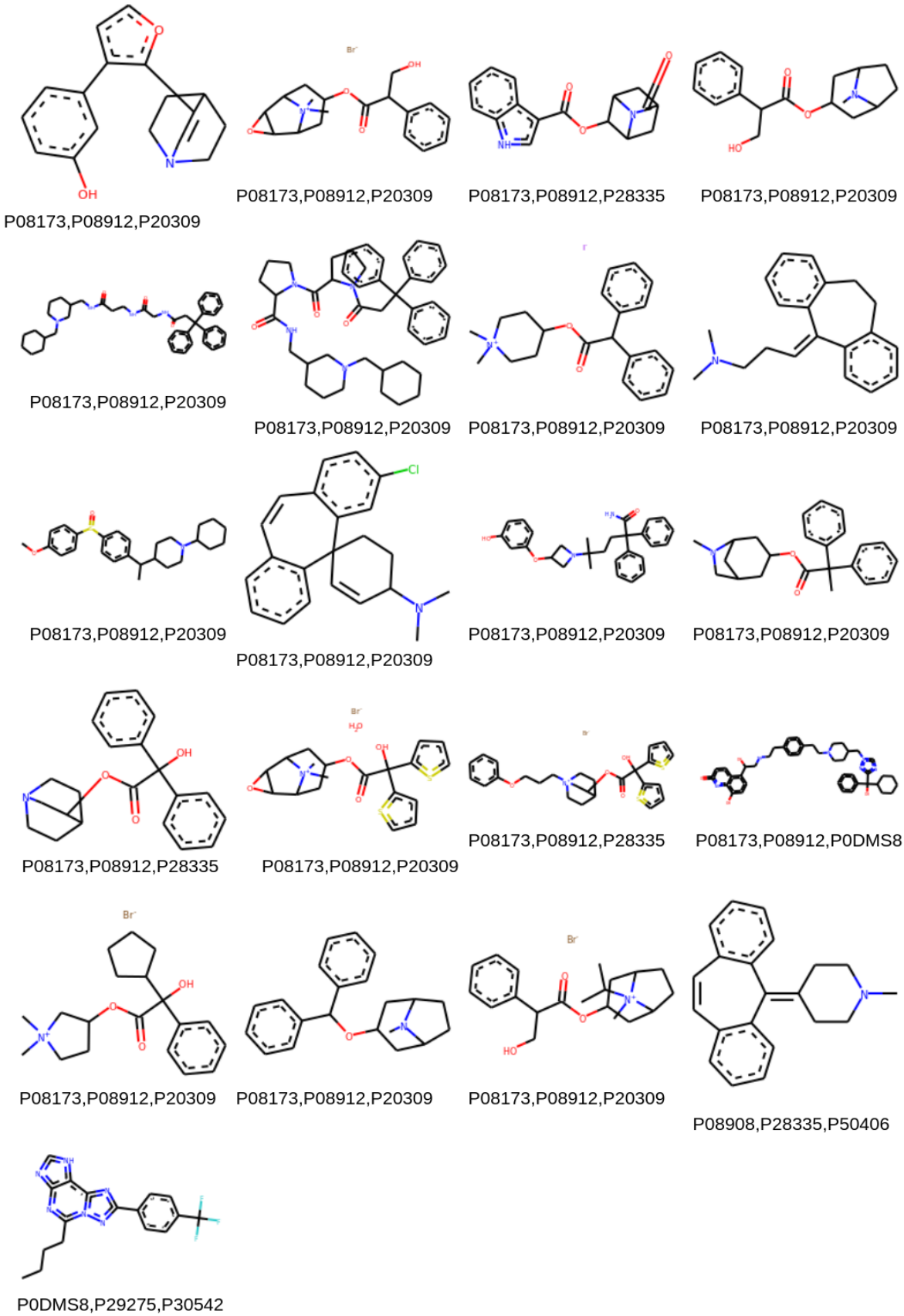


Figure 11 – Distribution of top potent ligands shared among at least 3 different GPCRs of all classes.

sistent with Lipinski's Rule of 5. Similarly, (MORPHY; RANKOVIC, 2006), evaluated the physicochemical properties for GPCRs ligands. They found that GPCRs ligands possess: median of 8 rotatable bonds, median molecular weight of 450 (mean=503), median logP value of 4.4 (mean=4.2) and a median for polar surface area of 67 $Å^2$. These common physicochemical properties of potent ligands illustrates some characteristics that should be considered during evaluation of datasets for screening new GPCRs drugs. Applying them in a curation step could reduce data to be evaluated, trimming time and computational power without losing potential ligands molecules. It is also important to note that these characteristics were agreeable with Lipinski's Rule of 5 and may indicate a bias in the original screening libraries applied in the discovery of these ligands. It is also important to note that some ligands were shared among GPCRs. This characteristic should be carefully studied during drug development in order to avoid off target activity.

4.3 Developing GPCR ligand predictors

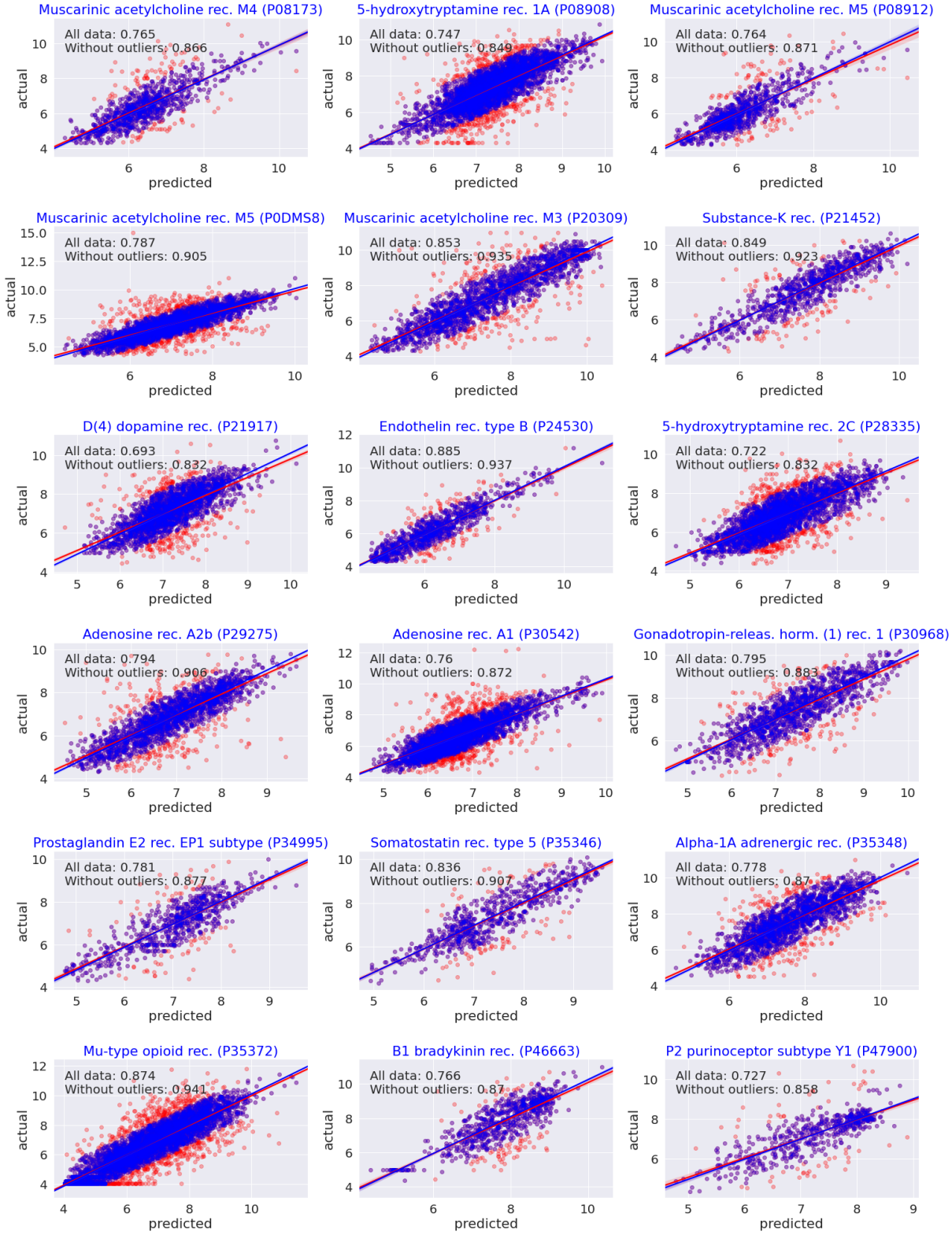
The first ML step in the development of our predictors was looking for the most meaningful features. We applied feature selection through 10-fold cross validation. For this task, we used a stepwise greedy feature selection algorithm. Table 7 present more details about our performance through 10-fold cross-validation before and after feature selection.

Final predictors models achieved Pearson's, Spearman's and Kendall's correlations of up to 0.89, 0.88 and 0.70 (median values, 0.78, 0.76, 0.59), respectively on 10-fold cross-validation (see Table 8, which are depicted as scatter plots (Figure 12)). We also assessed our models using Mean Square Error (MSE) to check how close our predictions were to the actual values. We reached a minimum of 0.24 and a maximum of 1.02. After 10% outlier removal, predictions improved substantially. For all receptors, the predictions reached a Pearson's correlation above 0.74 and considering the MSE values they decreased on average 40% for all receptors. We searched for outliers in common and found one, Clotrimazole, shared between four different receptors, D (4) dopamine receptor (P21917), 5-hydroxytryptamine receptor 2C (P28335), adenosine receptor A1 (P30542) and 5-hydroxytryptamine 9 receptor 6 (P50406). This finding reflects general properties of outliers, (outliers, meaning the 10% of molecules which presented the predictive value more distant from the regression curve) (see Figure A9 to Figure A22, Appendices) that tend to have less hydrogen bonds acceptors, hydrogen bonds donor, and negative ionisable atoms. We additionally found that other outliers tend to have also less positive ionisable atoms.

We also assessed the model under different cross validation schemes stratified 5- and 20-fold cross-validation (see Table 9) obtaining consistent results and demonstrating robustness and consistency of the models. Considering the supervised learning algorithms employed, 21 of the final models employed Random Forest and 10 Extra Trees, with the remaining 5 using

Table 7 – Predictors performance: first column represents performance using all graph-based signatures, second column, using all auxiliary features, third column, all graph signature combineed with all auxiliary features, and last column performance after feature selection (Pearson correlation coefficient on 10-fold cross-validation) (class A receptors are coloured in blue, class B in green, class C in red and class F in purple).

Receptor	Graph Signatures	Auxiliary Features	Graph + Auxiliary Features	Final
Muscarinic acetylcholine receptor M4-(P08173)	0.71	0.75	0.77	0.77
5-hydroxytryptamine receptor 1A-(P08908)	0.71	0.74	0.75	0.75
Muscarinic acetylcholine receptor M5-(P08912)	0.74	0.76	0.77	0.76
Muscarinic acetylcholine receptor M5-(P0DMS8)	0.75	0.78	0.79	0.79
Muscarinic acetylcholine receptor M3-(P20309)	0.82	0.85	0.85	0.85
Substance-K receptor-(P21452)	0.82	0.84	0.85	0.85
D(4) dopamine receptor-(P21917)	0.67	0.68	0.69	0.69
Endothelin receptor type B-(P24530)	0.86	0.88	0.88	0.89
5-hydroxytryptamine receptor 2C-(P28335)	0.69	0.71	0.72	0.72
Adenosine receptor A2b-(P29275)	0.77	0.79	0.79	0.79
Adenosine receptor A1-(P30542)	0.72	0.75	0.76	0.76
Gonadotropin-releasing hormone (type 1) receptor 1-(P30968)	0.78	0.79	0.79	0.80
Prostaglandin E2 rec EP1 sub-(P34995)	0.78	0.76	0.78	0.78
Somatostatin receptor type 5-(P35346)	0.83	0.83	0.83	0.84
Alpha-1A adrenergic receptor-(P35348)	0.74	0.77	0.78	0.78
Mu-type opioid receptor-(P35372)	0.84	0.87	0.87	0.87
B1 bradykinin receptor-(P46663)	0.74	0.76	0.75	0.77
P2 purinoceptor subtype Y1-(P47900)	0.72	0.73	0.73	0.73
Melatonin receptor type 1A-(P48039)	0.69	0.70	0.73	0.73
5-Hydroxytryptamine receptor 6-(P50406)	0.76	0.79	0.79	0.79
C-C chemokine receptor type 3-(P51677)	0.82	0.84	0.84	0.84
Hydroxycarboxylic acid receptor 2-(Q8TDS4)	0.67	0.67	0.68	0.67
G protein-coupled bile acid receptor 1-(Q8TDU6)	0.68	0.69	0.70	0.70
Mas-related G pro-coup rec X1-(Q96LB2)	0.72	0.42	0.53	0.69
Sphingosine 1-phosphate receptor 3-(Q99500)	0.74	0.78	0.78	0.78
Melanin-concentrating hormone receptors 1-(Q99705)	0.74	0.76	0.77	0.77
Sphingosine 1-phosphate receptor 5-(Q9H228)	0.83	0.86	0.86	0.86
G protein-coupled receptor 35-(Q9HC97)	0.80	0.84	0.84	0.84
Histamine H3 receptor-(Q9Y5N1)	0.75	0.78	0.79	0.79
Prostaglandin D2 receptor 2-(Q9Y5Y4)	0.72	0.71	0.74	0.74
Glucagon receptor-(P47871)	0.81	0.82	0.82	0.83
Calcitonin gene-related peptide type 1 receptor-(Q16602)	0.82	0.83	0.83	0.83
Extracellular calcium-sensing receptor-(P41180)	0.69	0.74	0.74	0.74
Metabotropic glutamate receptor 2-(Q14416)	0.81	0.84	0.84	0.84
Metabotropic glutamate receptor 4-(Q14833)	0.74	0.85	0.82	0.80
Smoothened homolog-(Q99835)	0.70	0.72	0.74	0.74


Muscarinic acetylcholine rec. M4 (P08173)
All data: 0.765
Without outliers: 0.866
5-hydroxytryptamine rec. 1A (P08908)
All data: 0.747
Without outliers: 0.849
Muscarinic acetylcholine rec. M5 (P08912)
All data: 0.764
Without outliers: 0.871
Muscarinic acetylcholine rec. M5 (P0DMS8)
All data: 0.787
Without outliers: 0.905
Muscarinic acetylcholine rec. M3 (P20309)
All data: 0.853
Without outliers: 0.935
Substance-K rec. (P21452)
All data: 0.849
Without outliers: 0.923
D(4) dopamine rec. (P21917)
All data: 0.693
Without outliers: 0.832
Endothelin rec. type B (P24530)
All data: 0.885
Without outliers: 0.937
5-hydroxytryptamine rec. 2C (P28335)
All data: 0.722
Without outliers: 0.832
Adenosine rec. A2b (P29275)
All data: 0.794
Without outliers: 0.906
Adenosine rec. A1 (P30542)
All data: 0.76
Without outliers: 0.872
Gonadotropin-releas. horm. (1) rec. 1 (P30968)
All data: 0.795
Without outliers: 0.883
Prostaglandin E2 rec. EP1 subtype (P34995)
All data: 0.781
Without outliers: 0.877
Somatostatin rec. type 5 (P35346)
All data: 0.836
Without outliers: 0.907
Alpha-1A adrenergic rec. (P35348)
All data: 0.778
Without outliers: 0.87
Mu-type opioid rec. (P35372)
All data: 0.874
Without outliers: 0.941
B1 bradykinin rec. (P46663)
All data: 0.766
Without outliers: 0.87
P2 purinoceptor subtype Y1 (P47900)
All data: 0.727
Without outliers: 0.858
actual
predicted

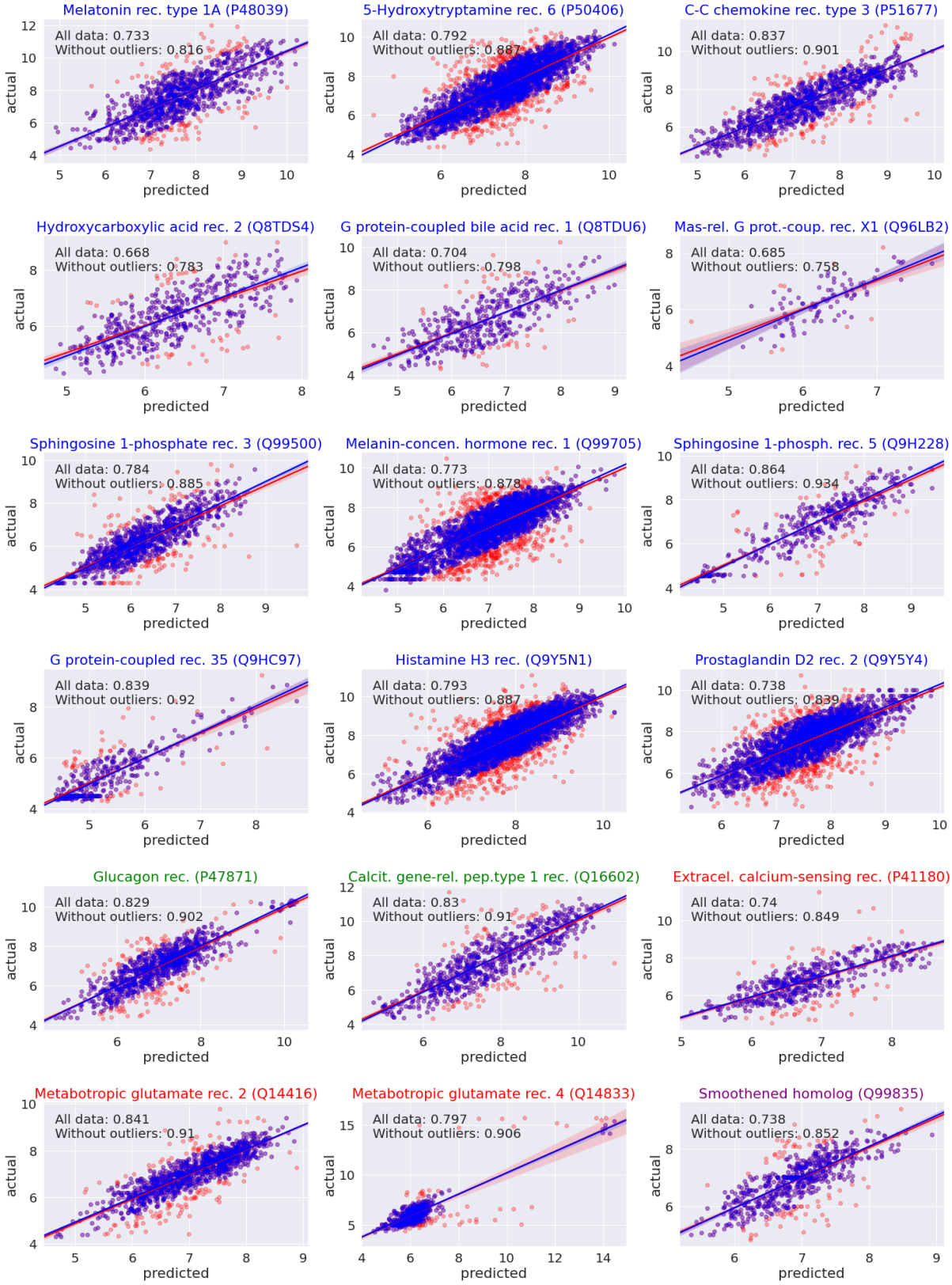


Figure 12 – Scatter plots, Regression analysis considering cross-validation schemes. Pearson's correlation coefficients are also shown in the top-left corner. These plots show the correlation between experimental (y-axis) and predicted values (x-axis).

Table 8 – Final Predictors' performance on 10-fold cross-validation. The values out of the parentheses mean all data, and the values in parentheses mean after 10% outlier removal (class A receptors are coloured in blue, class B in green, class C in red and class F in purple).

Receptor	Pearson	Spearman	Kendall	MSE
Muscarinic acetylcholine receptor M4-(P08173)	0.77(0.87)	0.75(0.83)	0.57(0.65)	0.59(0.29)
5-hydroxytryptamine receptor 1A-(P08908)	0.75 (0.85)	0.74(0.83)	0.56(0.64)	0.60(0.31)
Muscarinic acetylcholine receptor M5-(P08912)	0.76(0.87)	0.75(0.82)	0.56(0.63)	0.52(0.23)
Muscarinic acetylcholine receptor M5-(P0DMS8)	0.79(0.91)	0.80(0.90)	0.62(0.72)	0.55(0.23)
Muscarinic acetylcholine receptor M3-(P20309)	0.85(0.94)	0.85(0.94)	0.68(0.77)	0.70(0.31)
Substance-K receptor-(P21452)	0.85(0.92)	0.85(0.92)	0.67(0.75)	0.52(0.26)
D(4) dopamine receptor-(P21917)	0.69(0.83)	0.70(0.82)	0.52(0.63)	0.53(0.27)
Endothelin receptor type B-(P24530)	0.89(0.94)	0.86(0.92)	0.68(0.75)	0.34(0.18)
5-hydroxytryptamine receptor 2C-(P28335)	0.72(0.83)	0.71(0.82)	0.53(0.62)	0.52(0.29)
Adenosine receptor A2b-(P29275)	0.79(0.91)	0.80(0.90)	0.63(0.73)	0.46(0.20)
Adenosine receptor A1-(P30542)	0.76(0.87)	0.75(0.84)	0.57(0.65)	0.50(0.23)
Gonadotropin-releasing hormone (type 1) receptor 1-(P30968)	0.80(0.88)	0.79(0.88)	0.60(0.69)	0.52(0.29)
Prostaglandin E2 receptor EP1 subtype-(P34995)	0.78(0.88)	0.77(0.85)	0.59(0.69)	0.51(0.27)
Somatostatin receptor type 5-(P35346)	0.84(0.91)	0.82(0.89)	0.63(0.71)	0.45(0.25)
Alpha-1A adrenergic receptor-(P35348)	0.78(0.87)	0.78(0.87)	0.59(0.68)	0.63(0.35)
Mu-type opioid receptor-(P35372)	0.87(0.94)	0.88(0.94)	0.70(0.78)	0.62(0.29)
B1 bradykinin receptor-(P46663)	0.77(.87)	0.74(0.84)	0.55(0.64)	0.55(0.38)
P2 purinoceptor subtype Y1-(P47900)	0.73(0.86)	0.75(0.84)	0.56(0.65)	0.56(0.25)
Melatonin receptor type 1A-(P48039)	0.73(0.82)	0.73(0.80)	0.54(0.60)	1.02(0.62)
5-Hydroxytryptamine receptor 6-(P50406)	0.79(0.89)	0.78(0.88)	0.60(0.69)	0.53(0.28)
C-C chemokine receptor type 3-(P51677)	0.84(0.90)	0.83(.90)	0.65(0.71)	0.49(0.27)
Hydroxycarboxylic acid receptor 2-(Q8TDS4)	0.67(0.78)	0.67(0.78)	0.48(0.58)	0.52(0.32)
G protein-coupled bile acid receptor 1-(Q8TDU6)	0.70(0.80)	0.70(0.79)	0.50(0.58)	0.70(0.44)
Mas-related G protein-coupled receptor X1-(Q96LB2)	0.69(0.74)	0.49(0.70)	0.36(0.52)	0.47(0.25)
Sphingosine 1-phosphate receptor 3-(Q99500)	0.78(0.89)	0.77(0.87)	0.59(0.68)	0.43(0.21)
Melanin-concentrating hormone receptors 1-(Q99705)	0.77(0.88)	0.74(0.85)	0.57(0.66)	0.50(0.25)
Sphingosine 1-phosphate receptor 5-(Q9H228)	0.86(0.93)	0.85(0.91)	0.67(0.75)	0.44(0.21)
G protein-coupled receptor 35-(Q9HC97)	0.84(0.92)	0.76(0.81)	0.59(0.64)	0.24(0.11)
Histamine H3 receptor-(Q9Y5N1)	0.79(0.89)	0.79(0.87)	0.61(0.69)	0.47(0.23)
Prostaglandin D2 receptor 2-(Q9Y5Y4)	0.74(0.84)	0.73(0.82)	0.54(0.62)	0.54(0.30)
Glucagon receptor-(P47871)	0.83(.90)	0.80(0.87)	0.61(0.69)	0.38(0.20)
Calcitonin gene-related peptide type 1 receptor-(Q16602)	0.83(0.91)	0.83(0.91)	0.64(0.73)	0.83(0.43)
Extracellular calcium-sensing receptor-(P41180)	0.74(0.85)	0.74(0.83)	0.55(0.64)	0.50(0.25)
Metabotropic glutamate receptor 2-(Q14416)	0.84(0.91)	0.85(0.91)	0.67(0.73)	0.26(0.12)
Metabotropic glutamate receptor 4-(Q14833)	0.80(0.91)	0.58(0.67)	0.43(0.50)	1.07(0.24
Smoothened homolog-(Q99835)	0.74(0.85)	0.73(0.84)	0.55(0.65)	0.27(0.13)
Median	0.78(0.88)	0.76 (0.85)	0.59(0.66)	0.52(0.25)

XGBoost. Intriguingly, Gradient Boost was not selected for any of the receptors.

Moreover, we tested our models using low-redundancy independent blind test sets. Histograms were built to provide the distributions of the bioactivity labels for both the training and the low-redundancy independent blind tests datasets (Figure 13). The bioactivity is represented as -log[concentration] at the x-axis, concentration in molar value. And in the y-axis we have represented the frequency of molecules. Despite the low level of similarity between them, their bioactivity distributions were similar, and ranged from 4.5 to 9.5, and most of the molecules presented a bioactivity between 6 and 7. As already mentioned, (ZHANG et al., 2015) summarised ligand affinity data in solved GPCR structures, and found that Ki from ligands were generally values in the single-digit nM range, when we convert a single-digit nM (1nM, for example) using -log10(activity), we got a value of 9 (meaning our datasets covered active and inactive

Table 9 – Final Predictors cross-validation results using Pearson correlations on 5, 10 and 20-fold(class A receptors are coloured in blue, class B in green, class C in red and class F in purple).

Receptor	Final algorithm	5-Fold	10-Fold	20-Fold
Muscarinic acetylcholine receptor M4-(P08173)	Extra Trees	0.76	0.77	0.76
5-hydroxytryptamine receptor 1A-(P08908)	Random Forest	0.76	0.75	0.76
Muscarinic acetylcholine receptor M5-(P08912)	Extra Trees	0.77	0.76	0.77
Muscarinic acetylcholine receptor M5-(P0DMS8)	Random Forest	0.79	0.79	0.79
Muscarinic acetylcholine receptor M3-(P20309)	Extra Trees	0.85	0.85	0.85
Substance-K receptor-(P21452)	Random Forest	0.84	0.85	0.85
D(4) dopamine receptor-(P21917)	Random Forest	0.71	0.69	0.71
Endothelin receptor type B-(P24530)	Extra Trees	0.87	0.89	0.88
5-hydroxytryptamine receptor 2C-(P28335)	Random Forest	0.74	0.72	0.73
Adenosine receptor A2b-(P29275)	Random Forest	0.81	0.79	0.82
Adenosine receptor A1-(P30542)	Random Forest	0.77	0.76	0.77
Gonadotropin-releasing hormone (type 1) receptor 1-(P30968)	Random Forest	0.80	0.80	0.80
Prostaglandin E2 receptor EP1 subtype-(P34995)	Random Forest	0.79	0.78	0.80
Somatostatin receptor type 5-(P35346)	Extra Trees	0.85	0.84	0.84
Alpha-1A adrenergic receptor-(P35348)	Extra Trees	0.78	0.78	0.79
Mu-type opioid receptor-(P35372)	Extra Trees	0.87	0.87	0.87
B1 bradykinin receptor-(P46663)	XGBoost	0.75	0.77	0.76
P2 purinoceptor subtype Y1-(P47900)	Random Forest	0.71	0.73	0.74
Melatonin receptor type 1A-(P48039)	Random Forest	0.76	0.73	0.76
5-Hydroxytryptamine receptor 6-(P50406)	Random Forest	0.80	0.79	0.80
C-C chemokine receptor type 3-(P51677)	Random Forest	0.86	0.84	0.86
Hydroxycarboxylic acid receptor 2-(Q8TDS4)	XGBoost	0.67	0.67	0.67
G protein-coupled bile acid receptor 1-(Q8TDU6)	XGBoost	0.70	0.70	0.70
Mas-related G protein-coupled receptor X1-(Q96LB2)	XGBoost	0.68	0.69	0.69
Melanin-concentrating hormone receptors 1-(Q99705)	Random Forest	0.78	0.77	0.78
Smoothened homolog-(Q99835)	Random Forest	0.73	0.74	0.73
Sphingosine 1-phosphate receptor 5-(Q9H228)	Extra Trees	0.86	0.86	0.86
G protein-coupled receptor 35-(Q9HC97)	Random Forest	0.85	0.84	0.84
Histamine H3 receptor-(Q9Y5N1)	Extra Trees	0.80	0.79	0.80
Prostaglandin D2 receptor 2-(Q9Y5Y4)	Random Forest	0.76	0.74	0.75
Glucagon receptor-(P47871)	Extra Trees	0.84	0.83	0.83
Calcitonin gene-related peptide type 1 receptor-(Q16602)	Random Forest	0.84	0.83	0.84
Extracellular calcium-sensing receptor-(P41180)	XGBoost	0.75	0.74	0.75
Metabotropic glutamate receptor 2-(Q14416)	Random Forest	0.85	0.84	0.85
Metabotropic glutamate receptor 4-(Q14833)	Random Forest	0.74	0.80	0.72
Sphingosine 1-phosphate receptor 3-(Q99500)	Random Forest	0.76	0.78	0.77

small molecules). Considering the built histograms it is possible to visualise that even after the selection of the training set and test set trough clusterisation, activity (-log[Molar]) ranged from 4.5 to 9.5 (micro to nanomolar) in train and test distributions for most data sets. This step was done to avoid overfitting, a major problem in ML tasks. It is possible to verify that we achieved our goal in attempting to provide more generalised models, by virtue of the results obtained by our models.

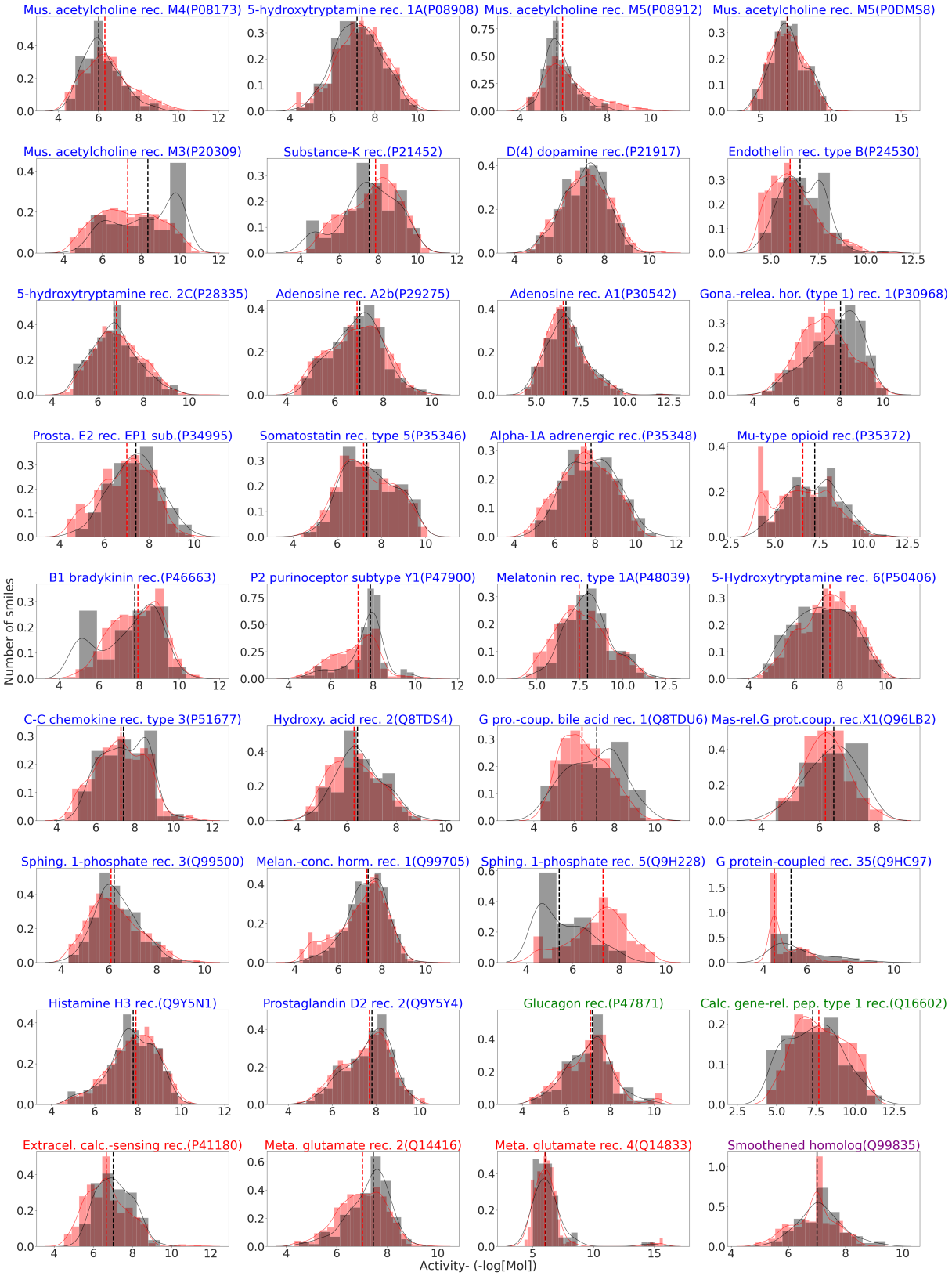


Figure 13 – Histograms considering molecular activity distribution for training and low-redundancy independent blind tests datasets. In x-axis the bioactivity is represented as -log[concentration], concentration in molar value. And in the y-axis, the frequency of molecules is represented. The histogram in red colour represents training and the histogram in grey colour represents testing datasets. The red line represents the median of the training set and the black line represents the median of the testintg set.

Predictive models for the 36 different GPCRs achieved Pearson's correlations up to 0.89, further demonstrating reliable predictive capabilities (see Table 10). Almost all final models had an increase in performance in the blind test after feature selection (28/36, the average increase in performance was 13.55%). The model developed to predict bioactivity for the "Mas-related G protein-coupled receptor X1" increased from 0.20 to 0.77. This receptor is an orphan GPCR, and we could use only 93 molecules. This result highlights the possibility of developing models even for targets with small datasets. Trying to understand better this fact, we generated a similarity matrix for the 93 molecules, using also Tanimoto index, comparing pairwise all the data (Figure 14). According to the similarity matrix, the dataset is very diverse in structures. We hypothesised that this data set characteristic together with the feature selection could explain the performance reached by this ML model. It was also intriguing to check that we had good performance not only for Family A GPCRs, but for all families evaluated. Family A is the more studied family, which reflects in the availability of data for the receptors of this family, and one could expect better results for them because of this fact.

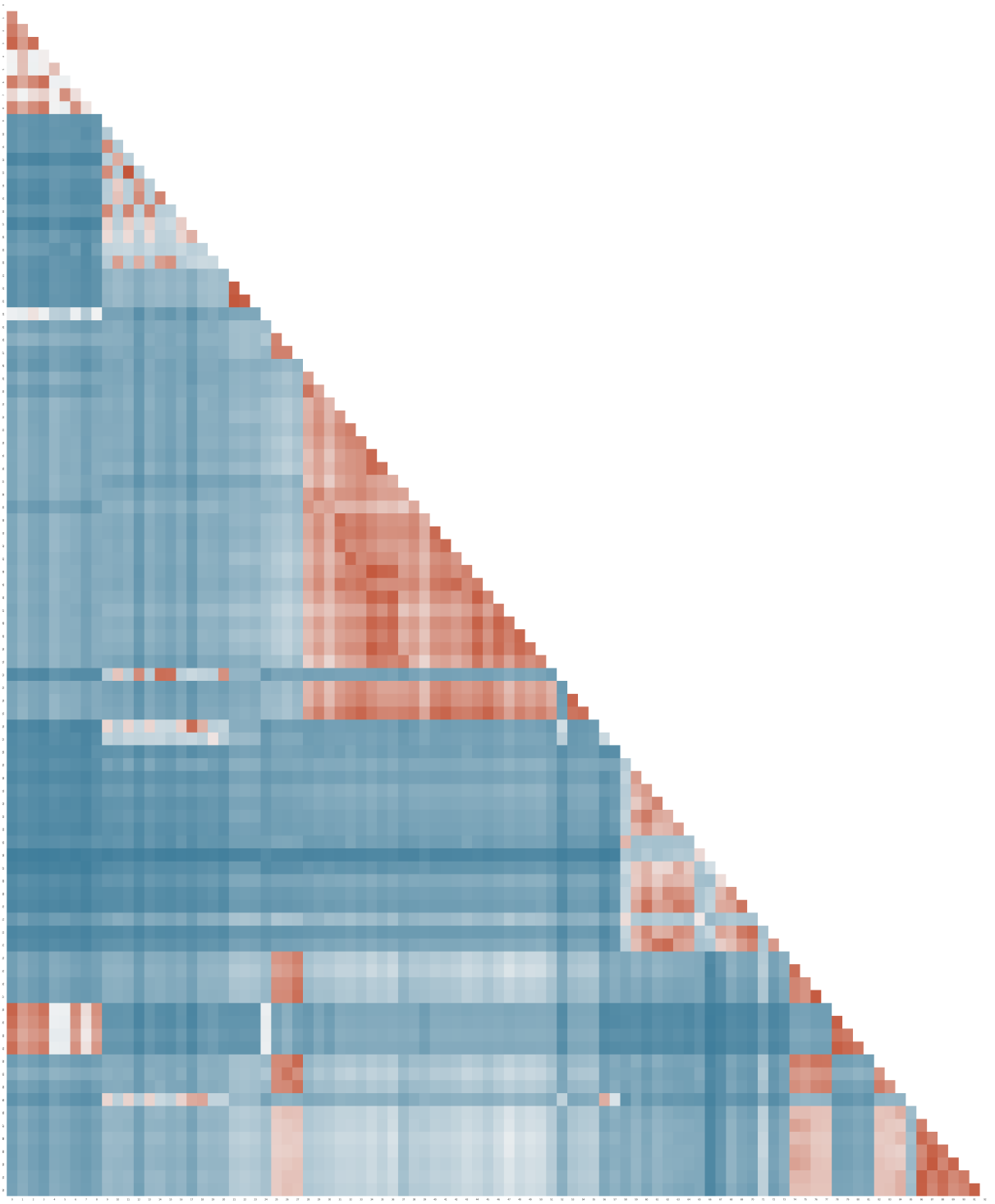


Figure 14 – Similarity matrix for the 93 molecules present in the Mas-related G protein-coupled receptor X1 data set. Each square means the Tanimoto score between two molecules. Blue colour means less similarity, red colour more similarity between the two compared smiles.

Table 10 – Blind test results, using all features and for the final models (class A receptors are coloured in blue, class B in green, class C in red and class F in purple).

Receptor	Using all Features (r)	Final Models (r)
Muscarinic acetylcholine receptor M4-(P08173)	0.59	0.59
5-hydroxytryptamine receptor 1A-(P08908)	0.63	0.67
Muscarinic acetylcholine receptor M5-(P08912)	0.63	0.69
Muscarinic acetylcholine receptor M5-(P0DMS8)	0.76	0.78
Muscarinic acetylcholine receptor M3-(P20309)	0.85	0.86
Substance-K receptor-(P21452)	0.83	0.86
D(4) dopamine receptor-(P21917)	0.54	0.62
Endothelin receptor type B-(P24530)	0.85	0.85
5-hydroxytryptamine receptor 2C-(P28335)	0.70	0.72
Adenosine receptor A2b-(P29275)	0.73	0.76
Adenosine receptor A1-(P30542)	0.71	0.72
Gonadotropin-releasing hormone (type 1) receptor 1-(P30968)	0.79	0.80
Prostaglandin E2 receptor EP1 subtype-(P34995)	0.64	0.71
Somatostatin receptor type 5-(P35346)	0.71	0.75
Alpha-1A adrenergic receptor-(P35348)	0.77	0.77
Mu-type opioid receptor-(P35372)	0.80	0.80
B1 bradykinin receptor-(P46663)	0.79	0.81
P2 purinoceptor subtype Y1-(P47900)	0.75	0.77
Melatonin receptor type 1A-(P48039)	0.60	0.68
5-Hydroxytryptamine receptor 6-(P50406)	0.77	0.78
C-C chemokine receptor type 3-(P51677)	0.82	0.84
Hydroxycarboxylic acid receptor 2-(Q8TDS4)	0.54	0.63
G protein-coupled bile acid receptor 1-(Q8TDU6)	0.76	0.82
Mas-related G protein-coupled receptor X1-(Q96LB2)	0.20	0.77
Sphingosine 1-phosphate receptor 3-(Q99500)	0.52	0.68
Melanin-concentrating hormone receptors 1-(Q99705)	0.67	0.68
Sphingosine 1-phosphate receptor 5-(Q9H228)	0.80	0.86
G protein-coupled receptor 35-(Q9HC97)	0.74	0.84
Histamine H3 receptor-(Q9Y5N1)	0.70	0.70
Prostaglandin D2 receptor 2-(Q9Y5Y4)	0.71	0.71
Glucagon receptor-(P47871)	0.59	0.69
Calcitonin gene-related peptide type 1 receptor-(Q16602)	0.76	0.80
Extracellular calcium-sensing receptor-(P41180)	0.65	0.73
Metabotropic glutamate receptor 2-(Q14416)	0.79	0.82
Metabotropic glutamate receptor 4-(Q14833)	0.89	0.89
Smoothened homolog-(Q99835)	0.84	0.84

4.4 Feature importance

Furthermore, we evaluated feature usage after feature selection per predictor. We used SHAP (SHapley Additive exPlanations) summary plots to illustrate feature importance (Figure 15 below and, Figure A23 to A26, Appendices, for the remaining receptors). SHAP assigns each feature an importance value for a particular prediction (LUNDBERG; LEE, 2017). The features are ordered by how much they influenced the model's prediction. The x-axis stands for the average of the absolute SHAP value of each feature. And the y-axis has the twenty more important features for each model. In the first plot, ML model for "Muscarinic acetylcholine receptor M4", it is possible to conclude that the most important feature for this model is SlogP_VSA5, followed

by SlogP_VSA6. These descriptors represent Subdivided logP Surface Areas, they are based on an approximate accessible van der Waals surface area calculation for each atom along with its contribution to logP (LABUTE, 2000). Checking SHAP bar plots for other models, it is possible to verify that these descriptors in addition to many graph-based signatures involving aromaticity, or hydrogen bond acceptors, stood out as influential features for the models. Through these findings it is possible to argue that the combination of conventional descriptors, such as logP, pKa, VSA, which are already established as relevant for drug transport or pharmacokinetics (PEARLMAN; SMITH, 1999), with graph based signatures can be of great support for drug screening. We postulate that this synergy between these two classes of features are due to the fact that these traditional descriptors are capable of describing whole molecule properties, and that the graph based signatures can account for substructural features. Furthermore, it is not surprising that we have various surface area descriptors selected as important, as was already cited by (LABUTE, 2000) that these descriptors weakly correlated with each other.

We also filtered the top ten features selected via the Forward Greedy Feature Selection approach for each of the machine learning models and calculated commonly used features. When considering only Class A receptors (Figure 16), two features were selected for 9 types of receptors (out of 31 class A receptors). The common features were topological polar surface area (TPSA) and also the presence of Bicyclic substructures on the molecule (fr_bicyclic), which is consistent with the most commonly found substructures in potent GPCR ligands. Five molecular surface area descriptors also stand out: SMR_VSA3, SMR_VSA7 (refers to Molecular refractivity combined with accessible van der Waals surface area contribution), SlogP_VSA2, SlogP_VSA8, SlogP_VSA3 (refers to Log of the aqueous solubility combined with accessible van der Waals surface area contribution), PEOE_VSA1 (refers to partial charges combined with accessible van der Waals surface area contribution). SMR_VSA3 was used on 8 receptor predictors, and the remaining on 7. Besides, 4 features encoding distance patterns were important, all involving pairs of aromatic atoms. These imply that the presence of aromatic rings on molecules are critical aspects of GPCR class A ligands. These selected features were consistent with considering all receptor types (Figure A27, Appendices). The results were consistent with SHAP bar plots.

4.5 Impacts of using different bioactivity measurements on performance

During the development of our models, we were concerned about mixing bioactivities as the done for previous methods (WU et al., 2018; WU et al., 2019; BURGGRAAFF et al., 2020; LIANG et al., 2019; KRUGER et al., 2014; ZIN; WILLIAMS; EKINS, 2020). We were worried, that even upon normalisation (-log[Molar]), that this could cause some noise in our predictions. Mainly because, for obtaining these different bioactivity values, different measurements are applied. Nevertheless, because of data availability, not mixing them were not feasible for many

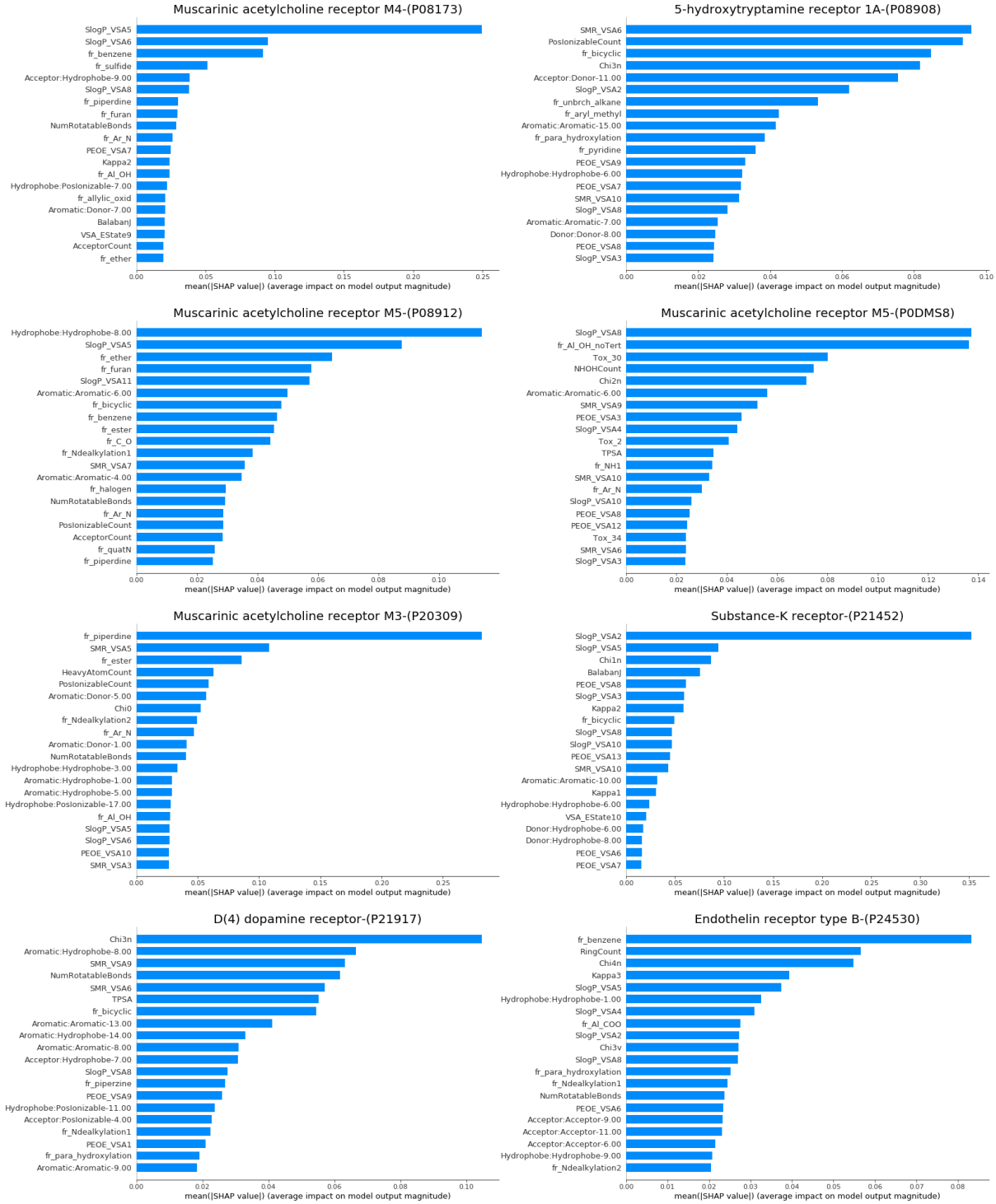


Figure 15 – SHAP bar plots - Feature importance for 8 class A receptors, the bars represent how strongly different input features affect the output of the respective model. Features are listed in descending order of importance.

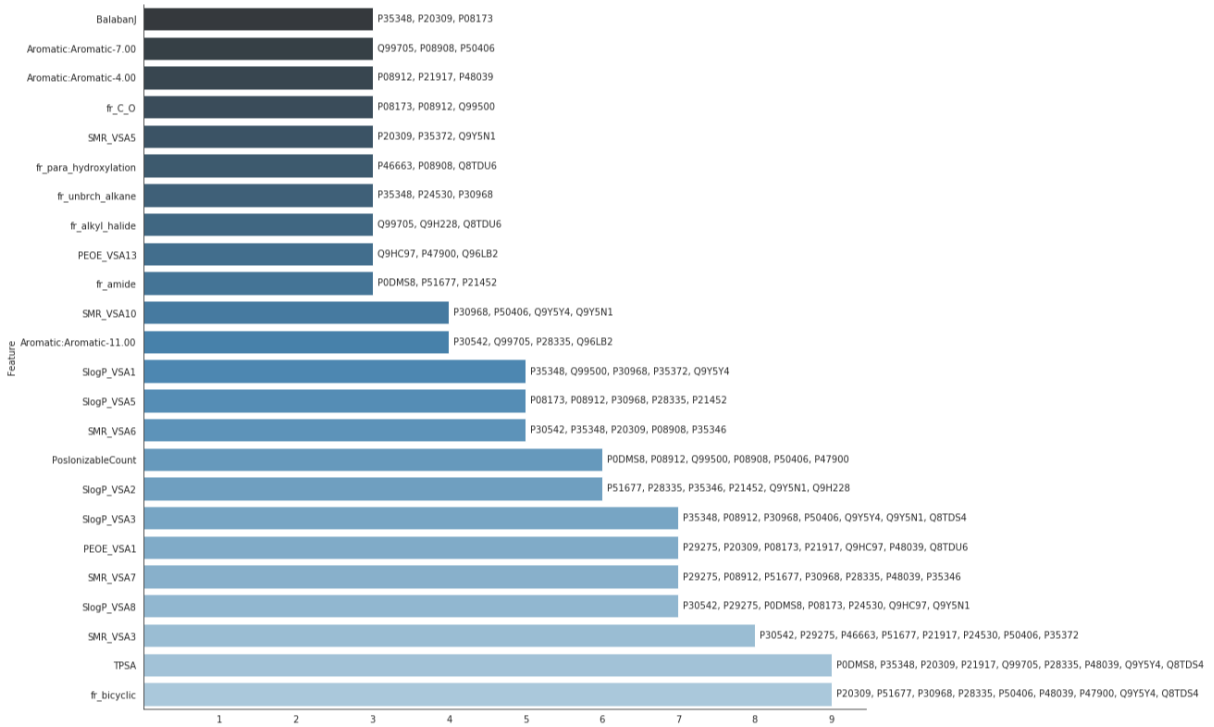


Figure 16 – Distribution of the top ten features selected via forward Greedy approach for Class A only receptors.

receptors, and adding to that, we wanted to compare the performance of our models against already published methods (WU et al., 2018), that mixed all the bioactivity. In order to clarify the impact of this aspect in our models, we have carried out a set of experiments training models using Ki+Kd values and testing them using IC50+EC50 (and vice-versa), depending on the number of molecules available in each case. The largest subset of activity types was assigned as the training set. Given data availability, this was performed for 13 different receptors, the remaining data did not enough data to test. The results showed a decrease in performance when predictors are trained and tested with different activity measures (Figure 17 for performance information and Figure 18 for molecular activity distribution). The worst performers on this scenario were Alpha-1A adrenergic receptor-(P35348), and 5-Hydroxytryptamine receptor 6-(P50406), Pearson's correlation of 0.06,and 0.09 respectively. These performance values indicate no correlation among training with Kd+Ki and testing EC50+IC50 for these receptors. It is a huge decrease of performance considering the final model performance 0.77, and 0.78, respectively. The best performers on this scenario were Melatonin receptor type 1A-(P48039), and Gonadotropin-releasing hormone (type 1) receptor 1-(P30968), Pearson's correlation of 0.50, and 0.54 respectively. These performance values indicate some correlation among training with Kd+Ki and testing EC50+IC50 for these receptors. Considering the final model performance of 0.68, and 0.80, respectively, the decreases in performance were approximately 30%. We assume that these decreases be due to inherent differences between bioactivity measurements, as already stated. And we acknowledge that, when possible, models should be trained and tested using only one type of bioactivity. Nonetheless, we also presume that these lower performances can be, as

Table 11 – Molecules for training pdCSM-GPCR that overlapped with datasets from WDL-RF. Second column represents number of molecules in pdCSMS-GPCR and the third the number in WDL-RF(class A receptors are coloured in blue, class B in green, class C in red and class F in purple).

Receptor	# pdCSMS-GPCR	# WDL-RF	Overlap (%)
5-hydroxytryptamine receptor 1A-(P08908)	3790	2294	43
Muscarinic acetylcholine receptor M5-(P08912)	959	369	27
Substance-K receptor-(P21452)	922	696	44
D(4) dopamine receptor-(P21917)	2335	1679	53
Endothelin receptor type B-(P24530)	987	1019	80
Adenosine receptor A1-(P30542)	3833	3016	50
Gonadotropin-releasing hormone (type 1) receptor 1-(P30968)	1373	1124	69
Prostaglandin E2 receptor EP1 subtype-(P34995)	741	236	17
Somatostatin receptor type 5-(P35346)	747	689	49
Alpha-1A adrenergic receptor-(P35348)	1898	1027	42
Mu-type opioid receptor-(P35372)	5275	3828	41
B1 bradykinin receptor-(P46663)	756	452	49
Melatonin receptor type 1A-(P48039)	1043	683	56
5-Hydroxytryptamine receptor 6-(P50406)	3044	1421	33
C-C chemokine receptor type 3-(P51677)	1131	781	61
Hydroxycarboxylic acid receptor 2-(Q8TDS4)	504	271	45
G protein-coupled bile acid receptor 1-(Q8TDU6)	443	1153	85
Sphingosine 1-phosphate receptor 3-(Q99500)	1088	317	20
Melanin-concentrating hormone receptors 1-(Q99705)	3721	2052	47
G protein-coupled receptor 35-(Q9HC97)	480	1579	76
Histamine H3 receptor-(Q9Y5N1)	3597	2092	45
Prostaglandin D2 receptor 2-(Q9Y5Y4)	2749	641	20
Glucagon receptor-(P47871)	1006	1129	72
Extracellular calcium-sensing receptor-(P41180)	535	940	70
Metabotropic glutamate receptor 2-(Q14416)	1168	1810	43
Smoothened homolog-(Q99835)	718	1523	67

well, due to biases in data set distributions (unbalanced data), and limited dataset sizes.

4.6 Comparative performance

In order to put our results into context, we compared the performance of our predictors with methods previously published (WU et al., 2018) (see Table 11) for information regarding overlap between molecules used in pdCSM-GPCR and WDL-RF).

The results are indicated in Figure 19, which shows that our predictors outperformed the alternative methods on almost all GPCR data sets, with statistically significant differences, except for the receptor Q14416 in which performances were very similar. The performances obtained in our models were comparable to the cross-validation performances, increasing our confidence in the method's generalisation capabilities.

We also plotted scatter plots for this step (Figure 20 for our models performance and Figure 21 for WDL-RF performance), and also a histogram which compares the activity outputs generated by the two servers, Figure 22. It was observed very high MSE measures for some WDL-RF models, which indicates high distance (discrepancy) between predicted and experimental

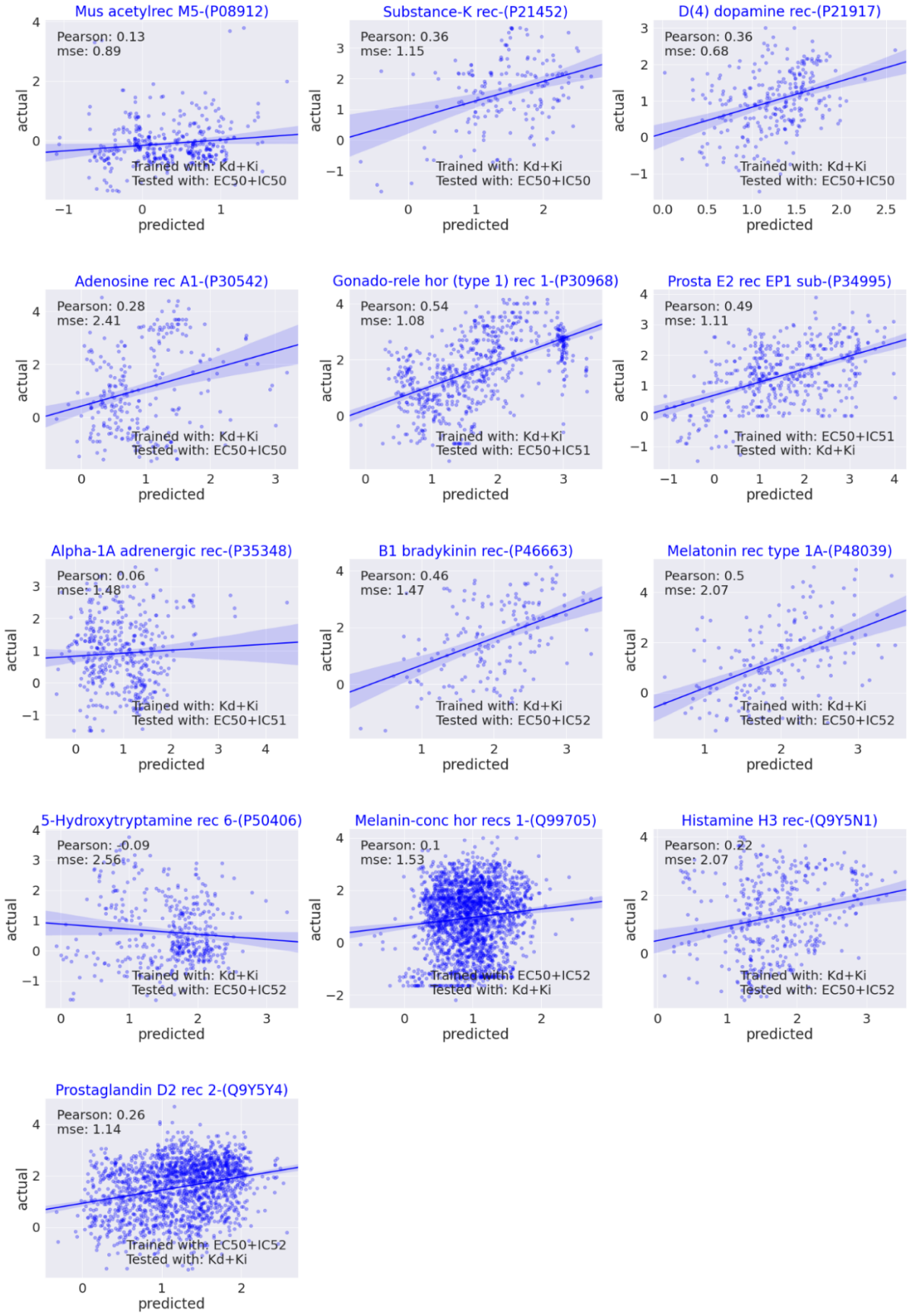


Figure 17 – Scatter plots - Regression analysis for training with a bioactivity and testing with another. Pearson's correlation coefficients and MSE are also shown in the top-left corner. The graphs show the correlation between experimental and predicted values.

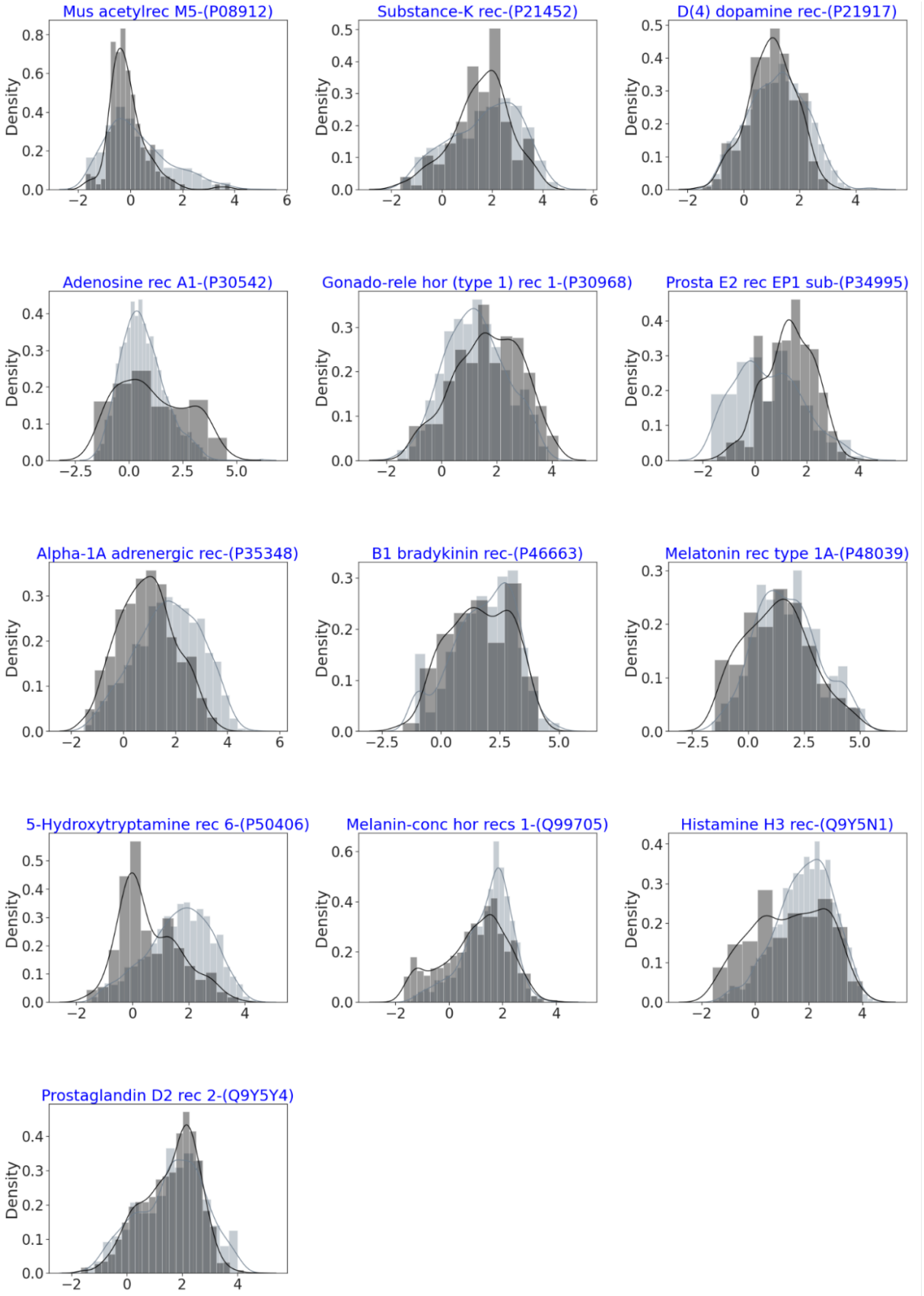


Figure 18 – Histograms considering molecular activity distribution for training with a bioactivity and testing with another. The histogram in light grey colour represents training and the histogram in dark grey colour represents testing datasets.

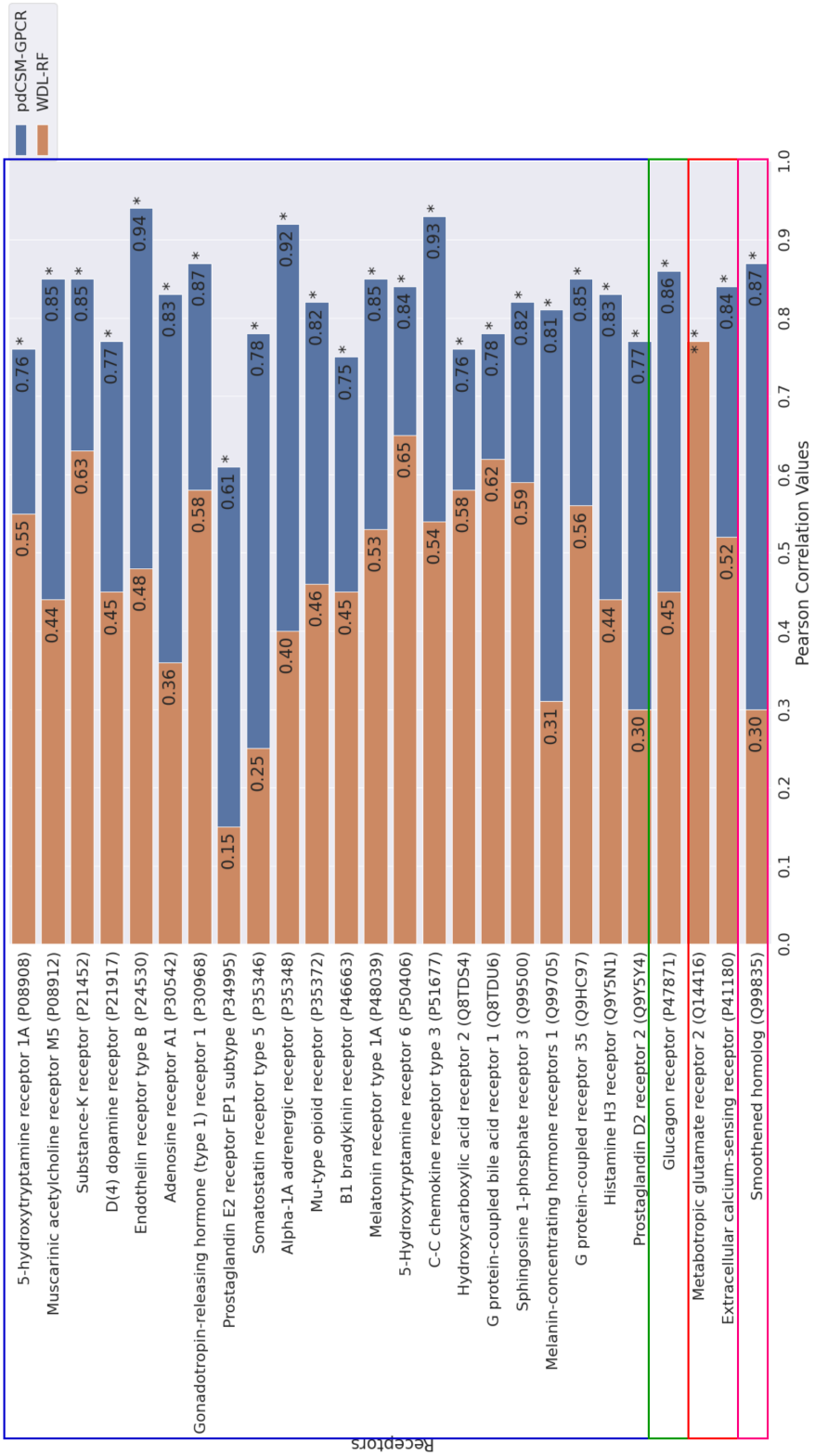


Figure 19 – Performance comparison between pdCSM-GPCR and (WU et al., 2018) (WDL-RF) through Pearson correlation. *Indicates that the pdCSM-GPCR significantly outperforms (p-value <0.001 using a Fisher's Z transformation). ** Pearson Correlation values were 0.75 for pdCSM-GPCR and 0.77 for WDL-RF.

values. We have also included Spearman and Kendall metrics, results obtained with them are consistent with our previous findings.

Interestingly, when carrying out this comparative analysis, we found that our models performed significantly better. This was unexpected because we adopted for these comparisons the datasets which were used for WDL-RF models training, and we expected at least values closer to the ones mentioned by them in the paper (WU et al., 2018). One possibility for this disparity, as already mentioned, might be the use of "control molecules". For doing this step, they use, for training, ligands that do not interact with the target GPCR. Because of that, they had hard set their bioactivity (-log10(activity in nM)) to -10, which could have increased their performance artificially. The control molecules were not available at the WDL-RF web server, and we could not use them for further testing. We, however, have also executed blind tests using a 'non-ligand' set. The small molecules for these 'non-ligand' sets were obtained through DUD-E (MYSINGER et al., 2012), a tool that generates decoys (non-ligands molecules) using active compounds. For this purpose, we used top potent ligands from our datasets. We added to our datasets 20% of decoys and the bioactivity of these were set to -1 (10 molar) (see Table 12 and Figure 23 to check the performance before and after adding decoys). The results we obtained demonstrated an increase in performance in 22 models out of 36 and for four occurred very little variation in performance. This demonstrates the robustness of our approach, but also shows how including decoys might overestimate performance of newly developed methods.

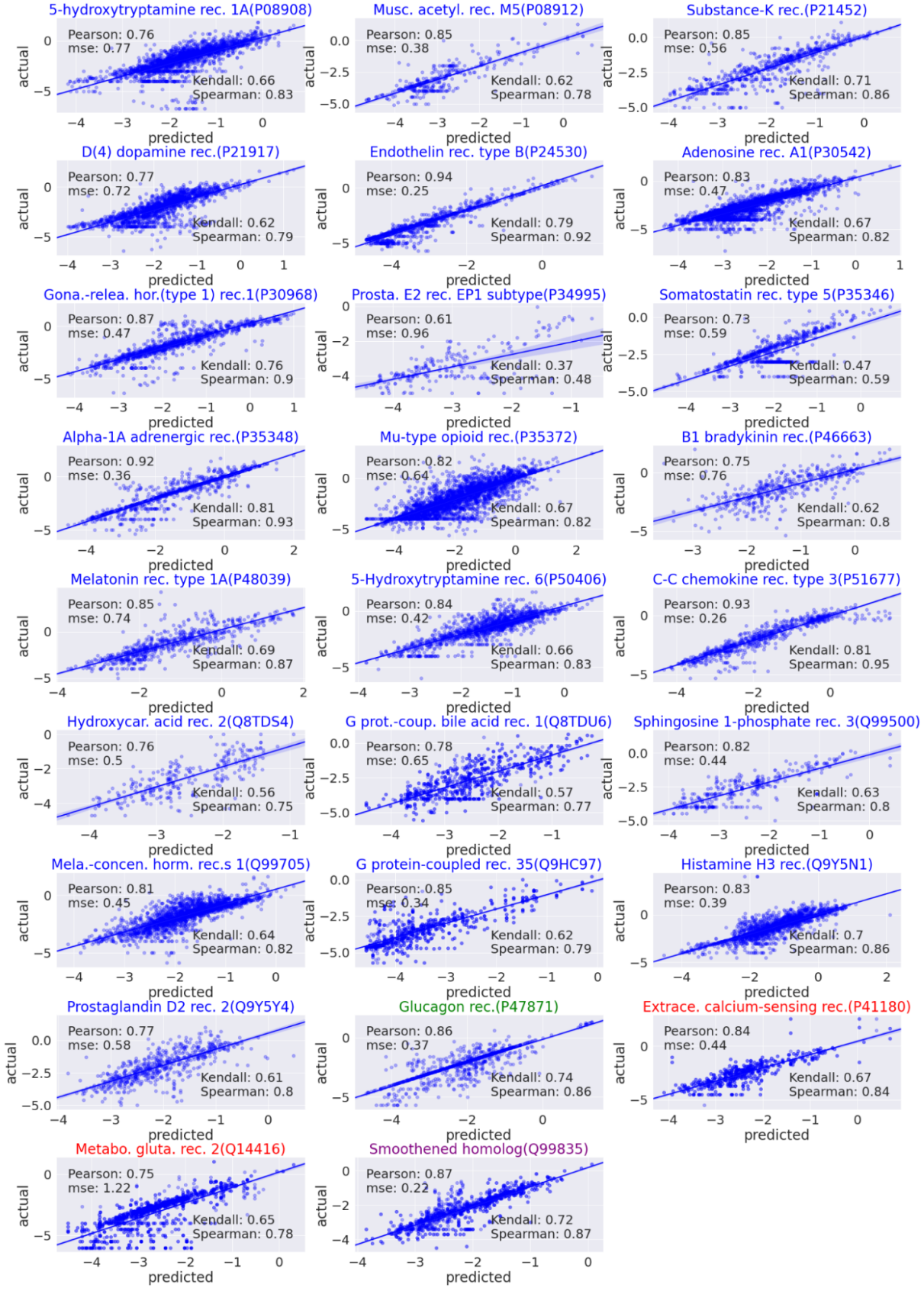


Figure 20 – Scatter plots - Regression analysis for pdCSM-GPCR when testing with WDL-RF datasets. Pearson's correlation coefficients and MSE are also shown in the top-left corner. The graphs show the correlation between experimental and predicted values.

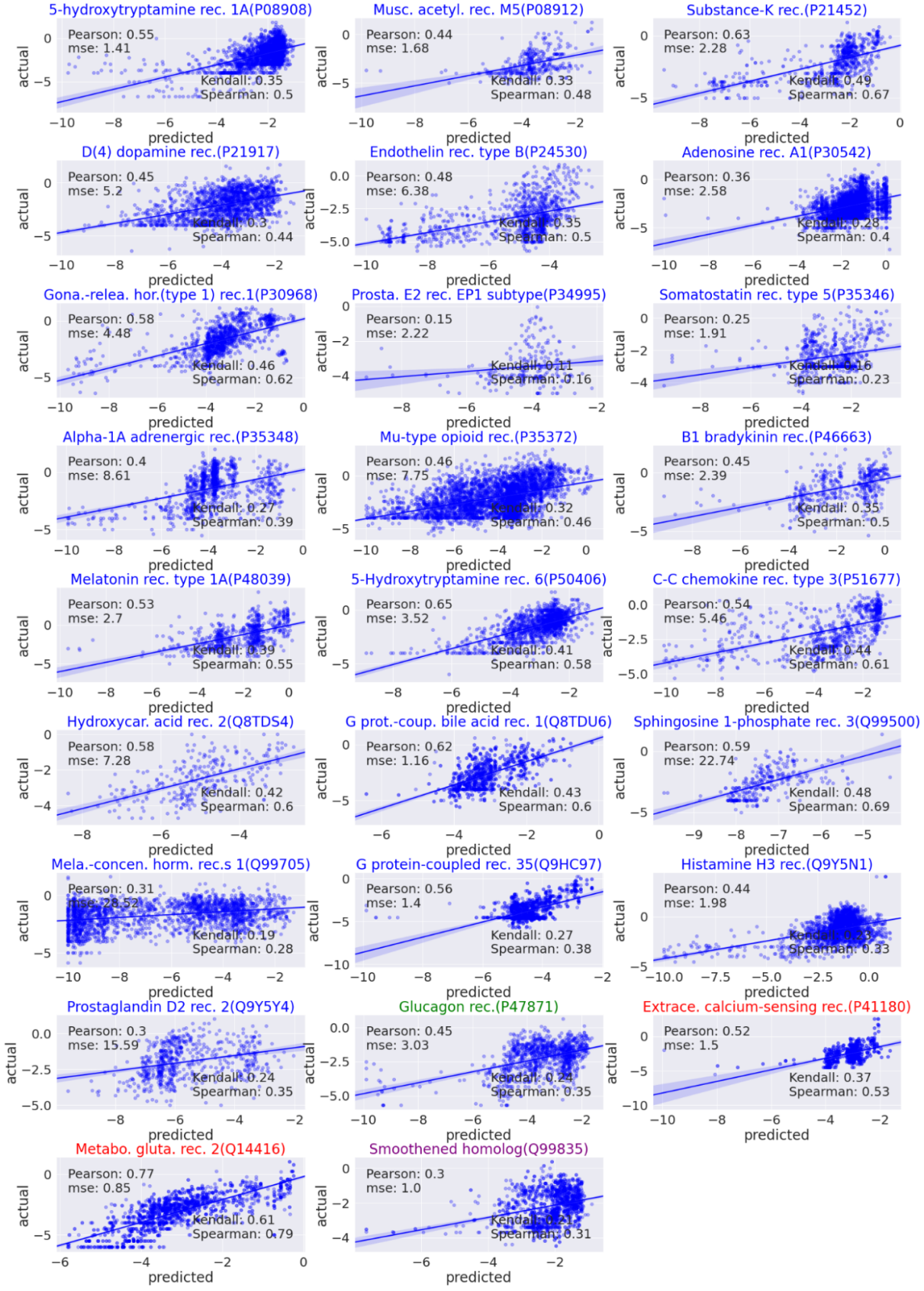


Figure 21 – Scatter plots - Regression analysis for WDL-RF when testing with WDL-RF datasets. Pearson's correlation coefficients and MSE are also shown in the top-left corner. The graphs show the correlation between experimental and predicted values.

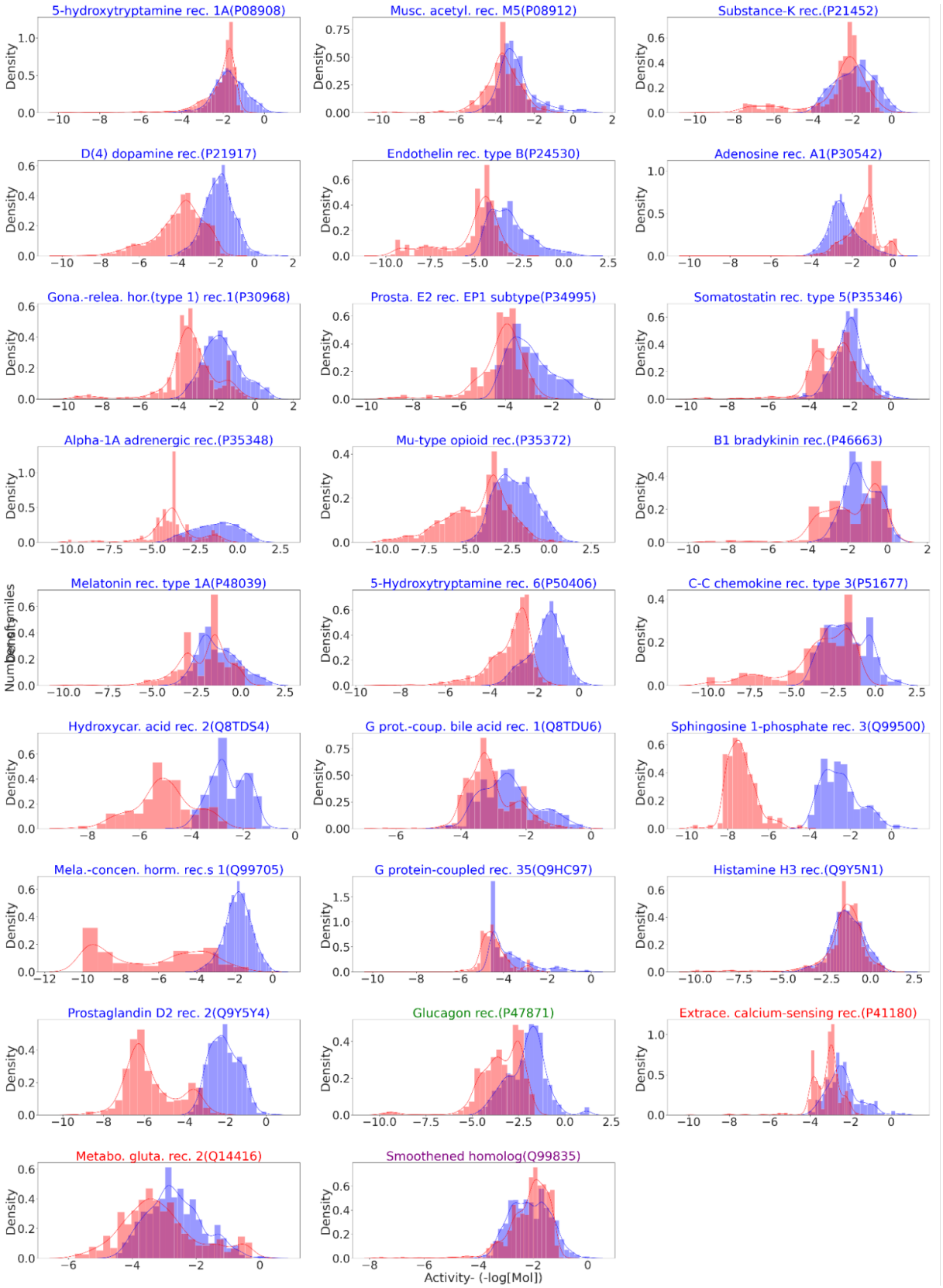


Figure 22 – Histogram - comparing the activity outputs predicted by the two servers: WDL-RF (in red), and pdCSM-GPCR (in blue).

Table 12 – Performance comparison between pdCSM-GPCR with and without decoys through Pearson's correlation. Green means that the model had a higher performance when using decoys (at least 0.01 higher or more) (class A receptors are coloured in blue, class B in green, class C in red and class F in purple).

Receptor	Without Decoys (r)	With Decoys (r)
Muscarinic acetylcholine receptor M4-(P08173)	0.59	0.66
5-hydroxytryptamine receptor 1A-(P08908)	0.67	0.93
Muscarinic acetylcholine receptor M5-(P08912)	0.69	0.72
Muscarinic acetylcholine receptor M5-(P0DMS8)	0.78	0.5
Muscarinic acetylcholine receptor M3-(P20309)	0.86	0.84
Substance-K receptor-(P21452)	0.86	0.92
D(4) dopamine receptor-(P21917)	0.62	0.79
Endothelin receptor type B-(P24530)	0.85	0.92
5-hydroxytryptamine receptor 2C-(P28335)	0.72	0.79
Adenosine receptor A2b-(P29275)	0.76	0.69
Adenosine receptor A1-(P30542)	0.72	0.59
Gonadotropin-releasing hormone (type 1) receptor 1-(P30968)	0.80	0.61
Prostaglandin E2 receptor EP1 subtype-(P34995)	0.71	0.9
Somatostatin receptor type 5-(P35346)	0.75	0.79
Alpha-1A adrenergic receptor-(P35348)	0.77	0.44
Mu-type opioid receptor-(P35372)	0.80	0.92
B1 bradykinin receptor-(P46663)	0.81	0.65
P2 purinoceptor subtype Y1-(P47900)	0.77	0.95
Melatonin receptor type 1A-(P48039)	0.68	0.44
5-Hydroxytryptamine receptor 6-(P50406)	0.78	0.88
C-C chemokine receptor type 3-(P51677)	0.84	0.98
Metabotropic glutamate receptor 4-(Q14833)	0.89	0.52
Calcitonin gene-related peptide type 1 receptor-(Q16602)	0.80	0.78
Hydroxycarboxylic acid receptor 2-(Q8TDS4)	0.63	0.9
G protein-coupled bile acid receptor 1-(Q8TDU6)	0.82	0.26
Mas-related G protein-coupled receptor X1-(Q96LB2)	0.77	0.54
Sphingosine 1-phosphate receptor 3-(Q99500)	0.68	0.78
Melanin-concentrating hormone receptors 1-(Q99705)	0.68	0.92
Sphingosine 1-phosphate receptor 5-(Q9H228)	0.86	0.81
G protein-coupled receptor 35-(Q9HC97)	0.84	0.84
Histamine H3 receptor-(Q9Y5N1)	0.70	0.89
Prostaglandin D2 receptor 2-(Q9Y5Y4)	0.71	0.97
Glucagon receptor-(P47871)	0.69	0.93
Extracellular calcium-sensing receptor-(P41180)	0.73	0.75
Metabotropic glutamate receptor 2-(Q14416)	0.82	0.88
Smoothened homolog-(Q99835)	0.84	0.89

As the last experiment, we submitted the decoys generated in the previous step to our servers. We then compared the predicted bioactivity for the decoys with the actual bioactivity of the potent molecules used originally for creating the decoys through DUD-E (MYSINGER et al., 2012). For this purpose, we generated the histograms of Figure A28 (Appendices). For most of the receptors, the decoys' bioactivity predictions distribution were very similar to the distribution of bioactivity among potent ligands. Through this finding, we postulate that some of our models don't have features that enable them to distinguish between decoys and potent ligands. This finding could be related to the fact that decoys are computed based on similar physical properties but with different chemical structures (topology), and the models don't have features capable of describing differences in topology among decoys and potent ligands (MYSINGER et al., 2012).

According to the results, there are three receptors in which the mentioned difference in distribution are very small, Metabotropic glutamate receptor 4-(Q14833), G protein-coupled receptor 35-(Q9HC97), G protein-coupled bile acid receptor 1-(Q8TDU6). The models from

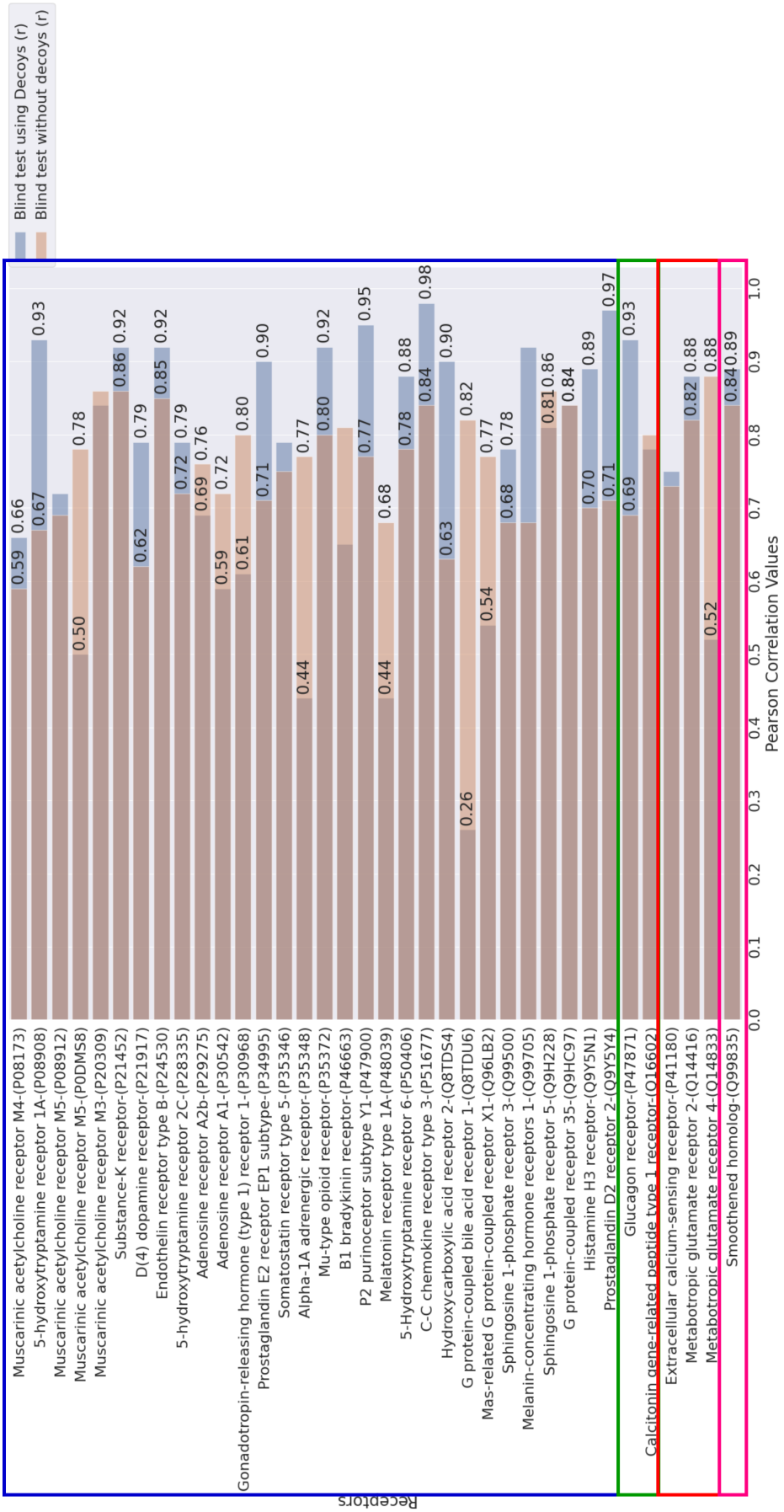


Figure 23 – Performance comparison between pdCSM-GPCR with and without decoys.

these receptors were the ones in which the performance using decoys decreased or kept the same (see Table 12). On these models, the decoys are being predicted as very potent ligands. The three receptors models in which the mentioned difference is higher, 5-hydroxytryptamine receptor 1A-(P08908), Histamine H3 receptor-(Q9Y5N1), Prostaglandin D2 receptor 2-(Q9Y5Y4), had an increase in performance when we added decoys (see Table 12). These models are capable of distinguishing decoys from ligands. According to this information, is evident the need of including features capable of distinguishing differences in topology. Another possibility is including docking approaches in the workflow. Throughout docking, this topological differences, potentially, would be identifiable.

4.7 pdCSM-GPCR Web Server Design and Implementation

The pdCSM-GPCR web server was designed to provide a user-friendly and quick web interface to predict bioactivity for GPCR ligands. The web server allows users to submit a single compound SMILES, or upload a list of them. Users can then choose which classes of receptor they want to generate bioactivity predictions for. When just a single compound is submitted, in addition to the bioactivity prediction result (in μMolar), the result's page also includes a molecule depiction and general molecular properties of the compound. When multiple compounds are submitted, the prediction results are displayed in an interactive table. All results can be downloaded as a comma-separated values (CSV) file (Figure 24).

Figure 24 – pdCSM-GPCR web server. (A) depicts the landing page for the resource. By clicking on “Prediction” (1) at the top menu, users are directed to the job submission page (B). Users have the options to either provide a set of molecules as a SMILES file (2) or individual molecules as a SMILES string (3). Users can select the type of prediction (4). After selecting the type of prediction, and once calculations are complete, users are redirected to a results page (C) where predictions for GPCRs bioactivity are presented (5). Users have the option to download the results (6).

5 Concluding Remarks and Conclusion

GPCRs are the most strenuously evaluated protein as a drug target. This happens, mainly because of their massive involvement in human pathophysiology. Despite the recent and significant progress in tools for GPCR drug discovery, these approaches are always challenged by their significant system resources or time requirements. Consequently, there will always be demand for better and faster computational models that can help identify potential drugs. Furthermore, to date, most tools available are restricted to one receptor type or presented limited performance (RASTELLI; PINZI, 2015).

Here, we presented a computational platform dedicated to GPCR ligand design, pdCSM-GPCR, comprehending 36 different GPCRs ligands belonging to four families (A, B1, C, and F). Our models are capable of quantitatively predicting ligand bioactivity for the most comprehensive set of GPCR types and classes (A, B, C, and F) to date. Our approach relies on graph-based signatures, which in other biological questions demonstrated success in outperforming earlier methods (PIRES; ASCHER; BLUNDELL, 2014; PIRES; ASCHER, 2016b; PIRES; BLUNDELL; ASCHER, 2015; PIRES et al., 2013; PIRES; ASCHER, 2016a; PIRES; ASCHER, 2017). We additionally relied on different auxiliary signatures capable of describing the general physicochemical properties of compounds. We provided, through this, thesis models capable of scalability, being up to handling large data sets, an important requirement for screening initiatives.

Furthermore, our predictors are all regression models with actual numeric outputs. This is of great importance during drug development due to the fact it allows the prioritisation of ligands. Through prioritisation, the process of finding new molecules can be faster and less costly (SCHUFFENHAUER; JACOBY, 2004). Another primal aspect of our tools is that they can guide repurposing opportunities within known drugs, and support screen compound databases for potential GPCR ligands.

The results obtained here, support the idea that the lack of elucidated structure for receptors is not a constraint for the development of ligand predictors. And, the same procedure could be used for the development of any other receptor which already had been screened for new ligands, such as kinases, which also composes a great family of proteins super important for human biochemistry (MANNING et al., 2002). According to (AHMED et al., 2021), ligand-based models have the clear advantage over target-based models, because they applicable to any target with at least some reasonable number of known ligands present.

One observed limitation during our study is that good models are inherently linked to the availability of good data sets. In the direction of guaranteeing it, we carefully curated all our data, removing all repeated molecules and also with incoherent results. For example, molecules

that are marked as positive and negative activity. Because of that, we ended up with a great reduction in our final dataset. We are aware that we could have used less stringent parameters for this step, for example, checked the incoherent results and removed just the one less frequent or used the average of bioactivity, but because of the great number of datasets, this was infeasible. Nonetheless, our final results demonstrated good reliability and even for targets with low number of ligands. For example, Mas-related G protein-coupled receptor X1 (Q96LB2), our model performed with a Pearson correlation of 0.69, on cross-validation 10-fold.

Another critical point was regarding the mixing up of bioactivities. We were mindful that this could cause some noise in our predictions. In order to clarify that, we ran some tests and observed that indeed a decrease in performance occurs when predictors are trained and tested with different activity measures, a practice that was deployed by us and others (WU et al., 2018; WU et al., 2019; BURGGRAAFF et al., 2020; LIANG et al., 2019; KRUGER et al., 2014; ZIN; WILLIAMS; EKINS, 2020), and applied in some databases to allow comparable measures between different types of bioactivities (OVERINGTON; AL-LAZIKANI; HOPKINS, 2006), such as the pChEMBL value ($-\log10$ (molar IC50, XC50, EC50, AC50, Ki, Kd or Potency)) (BENTO et al., 2013). We presume that this might be due to biases in data set distributions, limited dataset sizes, as well as inherent differences between bioactivity measurements. Because of this, we acknowledge that, when possible, models should be trained and tested using only one type of bioactivity.

Moreover, as overfitting is also considered a major limitation in machine learning tasks, we tried to avoid that through clustering molecules and selecting test and training sets with some degree of dissimilarity. This would show us the real predictive capabilities of our models when facing unrelated molecules to the training set. Nonetheless, it is important to point out that the activities (measured within $-\log10$[Molar]) usually ranged from 4.5 to 9.5 (micro to nanomolar) and were consistent between train and test distributions for most data sets. Through, this approach, we verified that our predictors are really capable of good performance on previously unseen data.

We compared our predictive model's performances with WDL-RF (WU et al., 2018). The comparison was done using the data sets provided by the authors while training their models available online. WDL-RF makes use of control molecules, molecules which do not interact with the target GPCR and because of it, they set their bioactivity ($-\log10$[nM]) to -10. Also, according to them, this step was done to obtain more robust regression models. The control molecules they used were not available at their web server. In order to check the influence of these control molecules, we used DUD-E (MYSINGER et al., 2012), a tool that generates decoys using active compounds. According to our results, we concluded that the use of decoys is not trustful, we postulate that it happens because arbitrarily for a considerable part of the data set the bioactivity is set to only one value, which inserts bias and induces overestimated performance.

We also improved our understanding of the molecular properties of GPCRs ligands. According to our findings, aromatic rings and nitrogen containing fragments were most common

in GPCR potent ligands across all classes. Our findings corroborate with (HORST et al., 2009). This study also used frequent substructure mining to analyse the structural features of GPCR ligands. They found that the best discriminating substructures (they are more frequent in active than inactive molecules) have a symmetrical organisation of lipophilicity (aliphatic carbon atoms) around a Nitrogen. (STRADER et al., 1988) identified, using mutagenesis, a negatively charged aspartic acid residue in transmembrane domain 3 of the β -adrenoceptor. This residue is found to form a salt bridge with the ligands' protonated amino group. The presence of the nitrogen containing fragments also could be correlated to the importance of hydrogen donors in the interactions with their target.

In future work, we would like to focus on developing more assertive predictors using more layers of knowledge, such as GPCRs structural data (including information of the binding sites of GPCRS) coupled with interaction fingerprints. These fingerprints are binary 1D representations that encode information of the 3D structure of protein–ligand complexes (VASS et al., 2016). Combining this information with machine learning methods, more relationships between bioactivity and features can be learned, creating support for the development of new ligands that can have a specific role in the signalling such as allostery or capable of inducing a specific effect in GPCRs (inducing one of the many conformations that exist in the large spectrum of GPCRs states). In addition, we would like to further expand our predictions, including information on absorption, distribution, metabolism, excretion and toxicity data specific for GPCRs, developing methods for optimising small molecules tailored for GPCRs. This task would have a transformative effect on the drug development process, reducing the high attrition rates that exists in drug development (WARING et al., 2015).

All models developed and described in this thesis were made freely available via easy-to-use web server. We strongly believe these represent invaluable resources that can help to accelerate GPCR's ligand discovery.

Bibliography

AHMED, M. et al. GPCR_LigandClassify.py a rigorous machine learning classifier for GPCR targeting compounds. *Scientific Reports*, Springer Science and Business Media LLC, v. 11, n. 1, maio 2021. Disponível em: <https://doi.org/10.1038/s41598-021-88939-5>. Citado na página 79.

ALEXANDER, S. P. H. et al. THE CONCISE GUIDE TO PHARMACOLOGY 2019/20: G protein-coupled receptors. *Br. J. Pharmacol.*, Wiley, v. 176 Suppl 1, n. S1, p. S21–S141, dez. 2019. Citado 2 vezes nas páginas 17 and 18.

BAEK, M. et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, American Association for the Advancement of Science (AAAS), v. 373, n. 6557, p. 871–876, ago. 2021. Disponível em: <https://doi.org/10.1126/science.abj8754>. Citado na página 22.

BALLESTEROS, J. A.; WEINSTEIN, H. [19] integrated methods for the construction of three-dimensional models and computational probing of structure-function relations in g protein-coupled receptors. In: *Methods in Neurosciences*. Elsevier, 1995. p. 366–428. Disponível em: <https://doi.org/10.1016/s1043-9471(05)80049-7>. Citado na página 20.

BASITH, S. et al. Exploring g protein-coupled receptors (GPCRs) ligand space via cheminformatics approaches: Impact on rational drug design. *Frontiers in Pharmacology*, Frontiers Media SA, v. 9, mar. 2018. Disponível em: <https://doi.org/10.3389/fphar.2018.00128>. Citado na página 27.

BENTO, A. P. et al. The ChEMBL bioactivity database: an update. *Nucleic Acids Research*, Oxford University Press (OUP), v. 42, n. D1, p. D1083–D1090, nov. 2013. Disponível em: <https://doi.org/10.1093/nar/gkt1031>. Citado na página 80.

BERTALOVITZ, A. C. et al. Frizzled-4 c-terminus distal to KTXXXW motif is essential for normal dishevelled recruitment and norrin-stimulated activation of Lef/Tcf-dependent transcriptional activation. *J. Mol. Signal.*, Ubiquity Press, Ltd., v. 11, p. 1, fev. 2016. Citado na página 20.

BHATTACHARYA, S. et al. Ligand-stabilized conformational states of human 2 adrenergic receptor: Insight into g-protein-coupled receptor activation. *Biophysical Journal*, Elsevier BV, v. 94, n. 6, p. 2027–2042, mar. 2008. Disponível em: <https://doi.org/10.1529/biophysj.107.117648>. Citado na página 28.

BHUDIA, N. et al. G protein-coupling of adhesion GPCRs ADGRE2/EMR2 and ADGRE5/CD97, and activation of G protein signalling by an anti-EMR2 antibody. *Sci. Rep.*, Springer Science and Business Media LLC, v. 10, n. 1, p. 1004, jan. 2020. Citado na página 17.

BILL, R. M. et al. Overcoming barriers to membrane protein structure determination. *Nature Biotechnology*, Springer Science and Business Media LLC, v. 29, n. 4, p. 335–340, abr. 2011. Disponível em: <https://doi.org/10.1038/nbt.1833>. Citado na página 21.

BOCK, A.; BERMUDEZ, M. Allosteric coupling and biased agonism in g protein-coupled receptors. *The FEBS Journal*, Wiley, v. 288, n. 8, p. 2513–2528, mar. 2021. Disponível em: <https://doi.org/10.1111/febs.15783>. Citado na página 28.

BORGELT, C.; MEINL, T.; BERTHOLD, M. MoSS. In: *Proceedings of the 1st international workshop on open source data mining frequent pattern mining implementations - OSDM '05*. ACM Press, 2005. Disponível em: <https://doi.org/10.1145/1133905.1133908>. Citado 2 vezes nas páginas 35 and 50.

BREIMAN, L. *Machine Learning*, Springer Science and Business Media LLC, v. 45, n. 1, p. 5–32, 2001. Disponível em: <https://doi.org/10.1023/a:1010933404324>. Citado na página 40.

BURGGRAAFF, L. et al. Quantitative prediction of selectivity between the a1 and a2a adenosine receptors. *Journal of Cheminformatics*, Springer Science and Business Media LLC, v. 12, n. 1, maio 2020. Disponível em: <https://doi.org/10.1186/s13321-020-00438-3>. Citado 3 vezes nas páginas 34, 64, and 80.

BUSATO, M.; GIORGETTI, A. Structural modeling of g-protein coupled receptors: An overview on automatic web-servers. *The International Journal of Biochemistry & Cell Biology*, Elsevier BV, v. 77, p. 264–274, ago. 2016. Disponível em: <https://doi.org/10.1016/j.biocel.2016.04.004>. Citado na página 21.

BUSHDID, C. et al. Agonists of g-protein-coupled odorant receptors are predicted from chemical features. *The Journal of Physical Chemistry Letters*, American Chemical Society (ACS), v. 9, n. 9, p. 2235–2240, abr. 2018. Disponível em: <https://doi.org/10.1021/acs.jpclett.8b00633>. Citado 2 vezes nas páginas 29 and 30.

BUTINA, D. Unsupervised data base clustering based on daylight's fingerprint and tanimoto similarity: A fast and automated way to cluster small and large data sets. *Journal of Chemical Information and Computer Sciences*, American Chemical Society (ACS), v. 39, n. 4, p. 747–750, jun. 1999. Disponível em: <https://doi.org/10.1021/ci9803381>. Citado na página 43.

CHEN, H. et al. A forward chemical genetic screen reveals gut microbiota metabolites that modulate host physiology. *Cell*, Elsevier BV, v. 177, n. 5, p. 1217–1231.e18, maio 2019. Disponível em: <https://doi.org/10.1016/j.cell.2019.03.036>. Citado na página 19.

CHEN, T.; GUESTRIN, C. XGBoost. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2016. Disponível em: <https://doi.org/10.1145/2939672.2939785>. Citado na página 40.

CHEREZOV, V. et al. High-resolution crystal structure of an engineered human 2 -adrenergic g protein–coupled receptor. *Science*, American Association for the Advancement of Science (AAAS), v. 318, n. 5854, p. 1258–1265, nov. 2007. Disponível em: <https://doi.org/10.1126/science.1150577>. Citado na página 21.

CHERFILS, J.; ZEGHOUF, M. Regulation of small GTPases by GEFs, GAPs, and GDIs. *Physiological Reviews*, American Physiological Society, v. 93, n. 1, p. 269–309, jan. 2013. Disponível em: <https://doi.org/10.1152/physrev.00003.2012>. Citado na página 22.

CHOTHIA, C.; LESK, A. The relation between the divergence of sequence and structure in proteins. *The EMBO Journal*, Wiley, v. 5, n. 4, p. 823–826, abr. 1986. Disponível em: <https://doi.org/10.1002/j.1460-2075.1986.tb04288.x>. Citado na página 21.

CHRISTOPHER, J. A. et al. Fragment and structure-based drug discovery for a class c GPCR: Discovery of the mGlu5 negative allosteric modulator HTL14242 (3-chloro-5-[6-(5-fluoropyridin-2-yl)pyrimidin-4-yl]benzonitrile). *Journal of Medicinal Chemistry*, American Chemical Society (ACS), v. 58, n. 16, p. 6653–6664, ago. 2015. Disponível em: <https://doi.org/10.1021/acs.jmedchem.5b00892>. Citado na página 28.

CHRISTOPOULOS, A. et al. G-protein-coupled receptor allosterism: the promise and the problem(s). *Biochemical Society Transactions*, Portland Press Ltd., v. 32, n. 5, p. 873–877, out. 2004. Disponível em: <https://doi.org/10.1042/bst0320873>. Citado na página 27.

CHUN, E. et al. Fusion partner toolchest for the stabilization and crystallization of g protein-coupled receptors. *Structure*, Elsevier BV, v. 20, n. 6, p. 967–976, jun. 2012. Disponível em: <https://doi.org/10.1016/j.str.2012.04.010>. Citado na página 21.

CIANCETTA, A. et al. Advances in computational techniques to study GPCR–ligand recognition. *Trends in Pharmacological Sciences*, Elsevier BV, v. 36, n. 12, p. 878–890, dez. 2015. Disponível em: <https://doi.org/10.1016/j.tips.2015.08.006>. Citado na página 19.

CONGREVE, M. et al. Impact of GPCR structures on drug discovery. *Cell*, Elsevier BV, v. 181, n. 1, p. 81–91, abr. 2020. Disponível em: <https://doi.org/10.1016/j.cell.2020.03.003>. Citado na página 24.

CONN, P. J.; CHRISTOPOULOS, A.; LINDSLEY, C. W. Allosteric modulators of GPCRs: a novel approach for the treatment of CNS disorders. *Nature Reviews Drug Discovery*, Springer Science and Business Media LLC, v. 8, n. 1, p. 41–54, jan. 2009. Disponível em: <https://doi.org/10.1038/nrd2760>. Citado na página 27.

COSTANZI, S. et al. Homology modeling of a class a GPCR in the inactive conformation: A quantitative analysis of the correlation between model/template sequence identity and model accuracy. *Journal of Molecular Graphics and Modelling*, Elsevier BV, v. 70, p. 140–152, nov. 2016. Disponível em: <https://doi.org/10.1016/j.jmgm.2016.10.004>. Citado na página 21.

DAVIES, M. N. et al. On the hierarchical classification of g protein-coupled receptors. *Bioinformatics*, Oxford University Press (OUP), v. 23, n. 23, p. 3113–3118, out. 2007. Disponível em: <https://doi.org/10.1093/bioinformatics/btm506>. Citado na página 46.

DIAZ, C.; ANGELLOZ-NICOUD, P.; PIHAN, E. Modeling and deorphanization of orphan GPCRs. In: *Methods in Molecular Biology*. Springer New York, 2017. p. 413–429. Disponível em: <https://doi.org/10.1007/978-1-4939-7465-8_21>. Citado na página 26.

DIGBY, G. J. et al. Some g protein heterotrimers physically dissociate in living cells. *Proceedings of the National Academy of Sciences*, Proceedings of the National Academy of Sciences, v. 103, n. 47, p. 17789–17794, nov. 2006. Disponível em: <https://doi.org/10.1073/pnas.0607116103>. Citado na página 22.

DORÉ, A. S. et al. Structure of class c GPCR metabotropic glutamate receptor 5 transmembrane domain. *Nature*, Springer Science and Business Media LLC, v. 511, n. 7511, p. 557–562, jul. 2014. Disponível em: <https://doi.org/10.1038/nature13396>. Citado na página 20.

DROR, R. O. et al. Pathway and mechanism of drug binding to g-protein-coupled receptors. *Proceedings of the National Academy of Sciences*, Proceedings of the National Academy of Sciences, v. 108, n. 32, p. 13118–13123, jul. 2011. Disponível em: <https://doi.org/10.1073/pnas.1104614108>. Citado na página 26.

ENGEL, T. Basic overview of chemoinformatics. *Journal of Chemical Information and Modeling*, American Chemical Society (ACS), v. 46, n. 6, p. 2267–2277, nov. 2006. Disponível em: <https://doi.org/10.1021/ci600234z>. Citado na página 30.

ERREY, J. C.; FIEZ-VANDAL, C. Production of membrane proteins in industry: The example of GPCRs. *Protein Expression and Purification*, Elsevier BV, v. 169, p. 105569, maio 2020. Disponível em: <https://doi.org/10.1016/j.pep.2020.105569>. Citado na página 21.

ESGUERRA, M. et al. GPCR-ModSim: A comprehensive web based solution for modeling g-protein coupled receptors. *Nucleic Acids Research*, Oxford University Press (OUP), v. 44, n. W1, p. W455–W462, maio 2016. Disponível em: <https://doi.org/10.1093/nar/gkw403>. Citado na página 21.

FISER, A.; DO, R. K. G.; ŠALI, A. Modeling of loops in protein structures. *Protein Science*, Wiley, v. 9, n. 9, p. 1753–1773, jan. 2000. Disponível em: <https://doi.org/10.1110/ps.9.9.1753>. Citado na página 21.

FOSTER, D. J.; CONN, P. J. Allosteric modulation of GPCRs: New insights and potential utility for treatment of schizophrenia and other CNS disorders. *Neuron*, Elsevier BV, v. 94, n. 3, p. 431–446, maio 2017. Disponível em: <https://doi.org/10.1016/j.neuron.2017.03.016>. Citado na página 27.

FOSTER, S. R. et al. Discovery of human signaling systems: Pairing peptides to g protein-coupled receptors. *Cell*, v. 179, n. 4, p. 895–908.e21, 2019. ISSN 0092-8674. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0092867419311262>. Citado 2 vezes nas páginas 16 and 27.

FREDRIKSSON, R. et al. The g-protein-coupled receptors in the human genome form five main families. phylogenetic analysis, paralogon groups, and fingerprints. *Molecular Pharmacology*, American Society for Pharmacology & Experimental Therapeutics (ASPET), v. 63, n. 6, p. 1256–1272, jun. 2003. Disponível em: <https://doi.org/10.1124/mol.63.6.1256>. Citado na página 17.

FRIEDMAN, J. H. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, Institute of Mathematical Statistics, v. 29, n. 5, out. 2001. Disponível em: <https://doi.org/10.1214/aos/1013203451>. Citado na página 40.

GARCÍA-NAFRÍA, J.; TATE, C. G. Structure determination of GPCRs: cryo-EM compared with x-ray crystallography. *Biochemical Society Transactions*, Portland Press Ltd., v. 49, n. 5, p. 2345–2355, set. 2021. Disponível em: <https://doi.org/10.1042/bst20210431>. Citado na página 21.

GAULTON, A. et al. The ChEMBL database in 2017. *Nucleic Acids Research*, Oxford University Press (OUP), v. 45, n. D1, p. D945–D954, nov. 2016. Disponível em: <https://doi.org/10.1093/nar/gkw1074>. Citado na página 29.

GEURTS, P.; ERNST, D.; WEHENKEL, L. Extremely randomized trees. *Machine Learning*, Springer Science and Business Media LLC, v. 63, n. 1, p. 3–42, mar. 2006. Disponível em: <https://doi.org/10.1007/s10994-006-6226-1>. Citado na página 40.

GHOSE, A. K.; VISWANADHAN, V. N.; WENDOLOSKI, J. J. A knowledge-based approach in designing combinatorial or medicinal chemistry libraries for drug discovery. 1. a qualitative and quantitative characterization of known drug databases. *Journal of Combinatorial*

Chemistry, American Chemical Society (ACS), v. 1, n. 1, p. 55–68, dez. 1998. Disponível em: <https://doi.org/10.1021/cc9800071>. Citado na página 50.

GILSON, M. K. et al. BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Research*, Oxford University Press (OUP), v. 44, n. D1, p. D1045–D1053, out. 2015. Disponível em: <https://doi.org/10.1093/nar/gkv1072>. Citado na página 29.

GUIDO, S.; MUELLER, A. C. *Introduction to machine learning with python*. Sebastopol, CA: O'Reilly Media, 2016. Citado na página 40.

HAUSER, A. S. et al. Trends in GPCR drug discovery: new agents, targets and indications. *Nature Reviews Drug Discovery*, Springer Science and Business Media LLC, v. 16, n. 12, p. 829–842, out. 2017. Disponível em: <https://doi.org/10.1038/nrd.2017.178>. Citado 4 vezes nas páginas 16, 25, 26, and 27.

HE, S.-B. et al. Predicting subtype selectivity for adenosine receptor ligands with three-dimensional biologically relevant spectrum (BRS-3d). *Scientific Reports*, Springer Science and Business Media LLC, v. 6, n. 1, nov. 2016. Disponível em: <https://doi.org/10.1038/srep36595>. Citado 2 vezes nas páginas 29 and 30.

HEIFETZ, A. et al. GPCR structure, function, drug discovery and crystallography: report from academia-industry international conference (UK royal society) chicheley hall, 1–2 september 2014. *Naunyn-Schmiedeberg's Archives of Pharmacology*, Springer Science and Business Media LLC, v. 388, n. 8, p. 883–903, mar. 2015. Disponível em: <https://doi.org/10.1007/s00210-015-1111-8>. Citado na página 29.

HEO, L.; FEIG, M. Multi-state modeling of g-protein coupled receptors at experimental accuracy. Cold Spring Harbor Laboratory, nov. 2021. Disponível em: <https://doi.org/10.1101/2021.11.26.470086>. Citado na página 22.

HERMANS, E. Biochemical and pharmacological control of the multiplicity of coupling at g-protein-coupled receptors. *Pharmacology & Therapeutics*, Elsevier BV, v. 99, n. 1, p. 25–44, jul. 2003. Disponível em: <https://doi.org/10.1016/s0163-7258(03)00051-2>. Citado na página 22.

HOLLENSTEIN, K. et al. Structure of class b GPCR corticotropin-releasing factor receptor 1. *Nature*, Springer Science and Business Media LLC, v. 499, n. 7459, p. 438–443, jul. 2013. Disponível em: <https://doi.org/10.1038/nature12357>. Citado na página 25.

HORST, E. van der et al. Substructure mining of GPCR ligands reveals activity-class specific functional groups in an unbiased manner. *Journal of Chemical Information and Modeling*, American Chemical Society (ACS), v. 49, n. 2, p. 348–360, fev. 2009. Disponível em: <https://doi.org/10.1021/ci8003896>. Citado 2 vezes nas páginas 50 and 81.

HU, B. et al. Three-dimensional biologically relevant spectrum (BRS-3d): Shape similarity profile based on PDB ligands as molecular descriptors. *Molecules*, MDPI AG, v. 21, n. 11, p. 1554, nov. 2016. Disponível em: <https://doi.org/10.3390/molecules21111554>. Citado 2 vezes nas páginas 29 and 30.

HUANG, X.-P. et al. Allosteric ligands for the pharmacologically dark receptors GPR68 and GPR65. *Nature*, Springer Science and Business Media LLC, v. 527, n. 7579, p. 477–483, nov. 2015. Disponível em: <https://doi.org/10.1038/nature15699>. Citado na página 29.

JABEEN, A.; RANGANATHAN, S. Applications of machine learning in GPCR bioactive ligand discovery. *Current Opinion in Structural Biology*, Elsevier BV, v. 55, p. 66–76, abr. 2019. Disponível em: <https://doi.org/10.1016/j.sbi.2019.03.022>. Citado 3 vezes nas páginas 10, 29, and 30.

JACOBSON, M. P. et al. A hierarchical approach to all-atom protein loop prediction. *Proteins: Structure, Function, and Bioinformatics*, Wiley, v. 55, n. 2, p. 351–367, mar. 2004. Disponível em: <https://doi.org/10.1002/prot.10613>. Citado na página 21.

JANG, J. W. et al. Novel scaffold identification of mGlu1 receptor negative allosteric modulators using a hierarchical virtual screening approach. *Chemical Biology & Drug Design*, Wiley, v. 87, n. 2, p. 239–256, out. 2015. Disponível em: <https://doi.org/10.1111/cbdd.12654>. Citado na página 30.

JIMÉNEZ-ROSÉS, M. et al. Combined docking and machine learning identifies key molecular determinants of ligand pharmacological activity on 2 adrenoceptor. Cold Spring Harbor Laboratory, mar. 2021. Disponível em: <https://doi.org/10.1101/2021.03.18.434755>. Citado na página 29.

JUMPER, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, Springer Science and Business Media LLC, v. 596, n. 7873, p. 583–589, jul. 2021. Disponível em: <https://doi.org/10.1038/s41586-021-03819-2>. Citado na página 22.

KATRITCH, V.; CHEREZOV, V.; STEVENS, R. C. Diversity and modularity of g protein-coupled receptor structures. *Trends in Pharmacological Sciences*, Elsevier BV, v. 33, n. 1, p. 17–27, jan. 2012. Disponível em: <https://doi.org/10.1016/j.tips.2011.09.003>. Citado na página 21.

KAZIUS, J.; MCGUIRE, R.; BURSI, R. Derivation and validation of toxicophores for mutagenicity prediction. *Journal of Medicinal Chemistry*, American Chemical Society (ACS), v. 48, n. 1, p. 312–320, dez. 2004. Disponível em: <https://doi.org/10.1021/jm040835a>. Citado na página 38.

KIM, S. et al. PubChem 2019 update: improved access to chemical data. *Nucleic Acids Research*, Oxford University Press (OUP), v. 47, n. D1, p. D1102–D1109, out. 2018. Disponível em: <https://doi.org/10.1093/nar/gky1033>. Citado 4 vezes nas páginas 29, 33, 34, and 46.

KOBILKA, B. K. Structural insights into adrenergic receptor function and pharmacology. *Trends in Pharmacological Sciences*, Elsevier BV, v. 32, n. 4, p. 213–218, abr. 2011. Disponível em: <https://doi.org/10.1016/j.tips.2011.02.005>. Citado na página 28.

KOBILKA, B. K.; DEUPI, X. Conformational complexity of g-protein-coupled receptors. *Trends in Pharmacological Sciences*, Elsevier BV, v. 28, n. 8, p. 397–406, ago. 2007. Disponível em: <https://doi.org/10.1016/j.tips.2007.06.003>. Citado na página 20.

KOOISTRA, A. J. et al. A structural chemogenomics analysis of aminergic GPCRs: lessons for histamine receptor ligand design. *British Journal of Pharmacology*, Wiley, v. 170, n. 1, p. 101–126, ago. 2013. Disponível em: <https://doi.org/10.1111/bph.12248>. Citado na página 50.

KOOISTRA, A. J. et al. GPCRdb in 2021: integrating GPCR sequence, structure and function. *Nucleic Acids Research*, Oxford University Press (OUP), v. 49, n. D1, p. D335–D343, dez. 2020. Disponível em: <https://doi.org/10.1093/nar/gkaa1080>. Citado na página 16.

KOUTSOUKAS, A. et al. Deep-learning: investigating deep neural networks hyper-parameters and comparison of performance to shallow methods for modeling bioactivity data. *Journal of Cheminformatics*, Springer Science and Business Media LLC, v. 9, n. 1, jun. 2017. Disponível em: <https://doi.org/10.1186/s13321-017-0226-y>. Citado 2 vezes nas páginas 29 and 30.

KRUGER, F. A. et al. PPDMs–a resource for mapping small molecule bioactivities from ChEMBL to pfam-a protein domains. *Bioinformatics*, Oxford University Press (OUP), v. 31, n. 5, p. 776–778, out. 2014. Disponível em: <https://doi.org/10.1093/bioinformatics/btu711>. Citado 3 vezes nas páginas 34, 64, and 80.

KRUSE, A. C. et al. Structure and dynamics of the m3 muscarinic acetylcholine receptor. *Nature*, Springer Science and Business Media LLC, v. 482, n. 7386, p. 552–556, fev. 2012. Disponível em: <https://doi.org/10.1038/nature10867>. Citado na página 26.

KRUSE, A. C. et al. Muscarinic acetylcholine receptors: novel opportunities for drug development. *Nature Reviews Drug Discovery*, Springer Science and Business Media LLC, v. 13, n. 7, p. 549–560, jun. 2014. Disponível em: <https://doi.org/10.1038/nrd4295>. Citado na página 27.

KUANG, Z.-K. et al. Predicting subtype selectivity of dopamine receptor ligands with three-dimensional biologically relevant spectrum. *Chemical Biology & Drug Design*, Wiley, v. 88, n. 6, p. 859–872, jul. 2016. Disponível em: <https://doi.org/10.1111/cbdd.12815>. Citado na página 29.

KURCZAB, R. et al. An algorithm to identify target-selective ligands – a case study of 5-HT7/5-HT1a receptor selectivity. *PLOS ONE*, Public Library of Science (PLoS), v. 11, n. 6, p. e0156986, jun. 2016. Disponível em: <https://doi.org/10.1371/journal.pone.0156986>. Citado 2 vezes nas páginas 29 and 30.

LABUTE, P. A widely applicable set of descriptors. *Journal of Molecular Graphics and Modelling*, Elsevier BV, v. 18, n. 4-5, p. 464–477, 2000. Disponível em: <https://doi.org/10.1016/s1093-3263(00)00068-1>. Citado na página 64.

LANGMEAD, C. J. et al. Identification of novel adenosine a2a receptor antagonists by virtual screening. *Journal of Medicinal Chemistry*, American Chemical Society (ACS), v. 55, n. 5, p. 1904–1909, fev. 2012. Disponível em: <https://doi.org/10.1021/jm201455y>. Citado na página 28.

LATORRACA, N. R.; VENKATAKRISHNAN, A. J.; DROR, R. O. GPCR dynamics: Structures in motion. *Chemical Reviews*, American Chemical Society (ACS), v. 117, n. 1, p. 139–155, set. 2016. Disponível em: <https://doi.org/10.1021/acs.chemrev.6b00177>. Citado na página 27.

LAUGWITZ, K. L. et al. The human thyrotropin receptor: a heptahelical receptor capable of stimulating members of all four g protein families. *Proceedings of the National Academy of Sciences*, Proceedings of the National Academy of Sciences, v. 93, n. 1, p. 116–120, jan. 1996. Disponível em: <https://doi.org/10.1073/pnas.93.1.116>. Citado na página 24.

LAUNAY, G. et al. Automatic modeling of mammalian olfactory receptors and docking of odorants. *Protein Engineering Design and Selection*, Oxford University Press (OUP), v. 25, n. 8, p. 377–386, jun. 2012. Disponível em: <https://doi.org/10.1093/protein/gzs037>. Citado na página 21.

LEBON, G.; WARNE, T.; TATE, C. G. Agonist-bound structures of g protein-coupled receptors. *Current Opinion in Structural Biology*, Elsevier BV, v. 22, n. 4, p. 482–490, ago. 2012. Disponível em: <https://doi.org/10.1016/j.sbi.2012.03.007>. Citado na página 27.

LEE, J.; FREDDOLINO, P. L.; ZHANG, Y. Ab initio protein structure prediction. In: *From Protein Structure to Function with Bioinformatics*. Springer Netherlands, 2017. p. 3–35. Disponível em: <https://doi.org/10.1007/978-94-024-1069-3_1>. Citado na página 21.

LEE, S. et al. How do branched detergents stabilize GPCRs in micelles? *Biochemistry*, American Chemical Society (ACS), v. 59, n. 23, p. 2125–2134, maio 2020. Disponível em: <https://doi.org/10.1021/acs.biochem.0c00183>. Citado na página 21.

LEFKOWITZ, R. J. A brief history of g-protein coupled receptors (nobel lecture). *Angewandte Chemie International Edition*, Wiley, v. 52, n. 25, p. 6366–6378, maio 2013. Disponível em: <https://doi.org/10.1002/anie.201301924>. Citado na página 22.

LIANG, L. et al. Bioactivity-explorer: a web application for interactive visualization and exploration of bioactivity data. *Journal of Cheminformatics*, Springer Science and Business Media LLC, v. 11, n. 1, jul. 2019. Disponível em: <https://doi.org/10.1186/s13321-019-0370-7>. Citado 3 vezes nas páginas 34, 64, and 80.

LUNDBERG, S. M.; LEE, S.-I. A unified approach to interpreting model predictions. In: GUYON, I. et al. (Ed.). *Advances in Neural Information Processing Systems 30*. Curran Associates, Inc., 2017. p. 4765–4774. Disponível em: <http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf>. Citado na página 63.

MAEHLE, A.-H. "receptive substances": John newport langley (1852–1925) and his path to a receptor theory of drug action. *Medical History*, Cambridge University Press (CUP), v. 48, n. 2, p. 153–174, abr. 2004. Disponível em: <https://doi.org/10.1017/s0025727300000090>. Citado na página 16.

MAEHLE, A.-H. A binding question: the evolution of the receptor concept. *Endeavour*, Elsevier BV, v. 33, n. 4, p. 135–140, dez. 2009. Disponível em: <https://doi.org/10.1016/j.endeavour.2009.09.001>. Citado na página 17.

MAGNANI, F. et al. Co-evolving stability and conformational homogeneity of the human adenosine a sub2a/sub receptor. *Proceedings of the National Academy of Sciences*, Proceedings of the National Academy of Sciences, v. 105, n. 31, p. 10744–10749, ago. 2008. Disponível em: <https://doi.org/10.1073/pnas.0804396105>. Citado na página 21.

MANNING, G. et al. The protein kinase complement of the human genome. *Science*, American Association for the Advancement of Science (AAAS), v. 298, n. 5600, p. 1912–1934, dez. 2002. Disponível em: <https://doi.org/10.1126/science.1075762>. Citado na página 79.

MASON, J. S. et al. New insights from structural biology into the druggability of g protein-coupled receptors. *Trends in Pharmacological Sciences*, Elsevier BV, v. 33, n. 5, p. 249–260, maio 2012. Disponível em: <https://doi.org/10.1016/j.tips.2012.02.005>. Citado na página 26.

MCCUDDEN, C. R. et al. G-protein signaling: back to the future. *Cellular and Molecular Life Sciences*, Springer Science and Business Media LLC, v. 62, n. 5, p. 551–577, mar. 2005. Disponível em: <https://doi.org/10.1007/s00018-004-4462-3>. Citado 2 vezes nas páginas 22 and 24.

MENDOZA, A. de; SEBÉ-PEDRÓS, A.; RUIZ-TRILLO, I. The evolution of the GPCR signaling system in eukaryotes: Modularity, conservation, and the transition to metazoan multicellularity. *Genome Biology and Evolution*, Oxford University Press (OUP), v. 6, n. 3, p. 606–619, fev. 2014. Disponível em: <https://doi.org/10.1093/gbe/evu038>. Citado 2 vezes nas páginas 16 and 23.

MILIC, D.; VEPRINTSEV, D. B. Large-scale production and protein engineering of g protein-coupled receptors for structural studies. *Frontiers in Pharmacology*, Frontiers Media SA, v. 6, mar. 2015. Disponível em: <https://doi.org/10.3389/fphar.2015.00066>. Citado na página 28.

MöHLER, H.; FRITSCHY, J. M.; RUDOLPH, U. A new benzodiazepine pharmacology. *Journal of Pharmacology and Experimental Therapeutics*, American Society for Pharmacology & Experimental Therapeutics (ASPET), v. 300, n. 1, p. 2–8, jan. 2002. Disponível em: <https://doi.org/10.1124/jpet.300.1.2>. Citado na página 27.

MORPHY, R.; RANKOVIC, Z. The physicochemical challenges of designing multiple ligands. *Journal of Medicinal Chemistry*, American Chemical Society (ACS), v. 49, n. 16, p. 4961–4970, jul. 2006. Disponível em: <https://doi.org/10.1021/jm0603015>. Citado na página 53.

MUNK, C. et al. Integrating structural and mutagenesis data to elucidate GPCR ligand binding. *Current Opinion in Pharmacology*, Elsevier BV, v. 30, p. 51–58, out. 2016. Disponível em: <https://doi.org/10.1016/j.coph.2016.07.003>. Citado na página 25.

MYSINGER, M. M. et al. Directory of useful decoys, enhanced (DUD-e): Better ligands and decoys for better benchmarking. *Journal of Medicinal Chemistry*, American Chemical Society (ACS), v. 55, n. 14, p. 6582–6594, jul. 2012. Disponível em: <https://doi.org/10.1021/jm300687e>. Citado 4 vezes nas páginas 45, 71, 75, and 80.

NAGI, K.; ONARAN, H. O. Biased agonism at g protein-coupled receptors. *Cellular Signalling*, Elsevier BV, v. 83, p. 109981, jul. 2021. Disponível em: <https://doi.org/10.1016/j.cellsig.2021.109981>. Citado na página 28.

NEW, D. C.; WONG, Y. H. Molecular mechanisms mediating the g protein-coupled receptor regulation of cell cycle progression. *Journal of Molecular Signaling*, Ubiquity Press, Ltd., v. 2, p. 2, fev. 2007. Disponível em: <https://doi.org/10.1186/1750-2187-2-2>. Citado na página 16.

NGO, T. et al. Identifying ligands at orphan GPCRs: current status using structure-based approaches. *British Journal of Pharmacology*, Wiley, v. 173, n. 20, p. 2934–2951, mar. 2016. Disponível em: <https://doi.org/10.1111/bph.13452>. Citado na página 28.

NICOLI, A. et al. Classification model for the second extracellular loop of class a GPCRs. *Journal of Chemical Information and Modeling*, American Chemical Society (ACS), v. 62, n. 3, p. 511–522, fev. 2022. Disponível em: <https://doi.org/10.1021/acs.jcim.1c01056>. Citado 2 vezes nas páginas 22 and 26.

NIIMURA, Y. Evolutionary dynamics of olfactory receptor genes in chordates: interaction between environments and genomic contents. *Human Genomics*, Springer Science and Business Media LLC, v. 4, n. 2, p. 107, 2009. Disponível em: <https://doi.org/10.1186/1479-7364-4-2-107>. Citado na página 17.

NISWENDER, C. M.; CONN, P. J. Metabotropic glutamate receptors: Physiology, pharmacology, and disease. *Annual Review of Pharmacology and Toxicology*, Annual Reviews, v. 50, n. 1, p. 295–322, fev. 2010. Disponível em: <https://doi.org/10.1146/annurev.pharmtox.011008.145533>. Citado na página 18.

OVERINGTON, J. P.; AL-LAZIKANI, B.; HOPKINS, A. L. How many drug targets are there? *Nature Reviews Drug Discovery*, Springer Science and Business Media LLC, v. 5, n. 12, p. 993–996, dez. 2006. Disponível em: <https://doi.org/10.1038/nrd2199>. Citado na página 80.

PALCZEWSKI, K. et al. Crystal structure of rhodopsin: A g protein-coupled receptor. *Science*, American Association for the Advancement of Science (AAAS), v. 289, n. 5480, p. 739–745, ago. 2000. Disponível em: <https://doi.org/10.1126/science.289.5480.739>. Citado na página 20.

PÁNDY-SZEKERES, G. et al. GPCRdb in 2018: adding GPCR structure models and ligands. *Nucleic Acids Research*, Oxford University Press (OUP), v. 46, n. D1, p. D440–D446, nov. 2017. Disponível em: <https://doi.org/10.1093/nar/gkx1109>. Citado 2 vezes nas páginas 28 and 29.

PARKER, J. L.; NEWSTEAD, S. Current trends in -helical membrane protein crystallization: An update. *Protein Science*, Wiley, v. 21, n. 9, p. 1358–1365, ago. 2012. Disponível em: <https://doi.org/10.1002/pro.2122>. Citado na página 21.

PEARLMAN, R. S.; SMITH, K. M. Metric validation and the receptor-relevant subspace concept. *Journal of Chemical Information and Computer Sciences*, American Chemical Society (ACS), v. 39, n. 1, p. 28–35, jan. 1999. Disponível em: <https://doi.org/10.1021/ci980137x>. Citado na página 64.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.*, JMLR.org, v. 12, n. null, p. 2825–2830, nov 2011. ISSN 1532-4435. Citado na página 42.

PIN, J.-P. et al. Allosteric functioning of dimeric class C g-protein-coupled receptors. *FEBS J.*, Wiley, v. 272, n. 12, p. 2947–2955, jun. 2005. Citado na página 17.

PIRES, D. E.; ASCHER, D. B. CSM-lig: a web server for assessing and comparing protein–small molecule affinities. *Nucleic Acids Research*, Oxford University Press (OUP), v. 44, n. W1, p. W557–W561, maio 2016. Disponível em: <https://doi.org/10.1093/nar/gkw390>. Citado 4 vezes nas páginas 30, 31, 38, and 79.

PIRES, D. E.; ASCHER, D. B. mCSM-AB: a web server for predicting antibody–antigen affinity changes upon mutation with graph-based signatures. *Nucleic Acids Research*, Oxford University Press (OUP), v. 44, n. W1, p. W469–W473, maio 2016. Disponível em: <https://doi.org/10.1093/nar/gkw458>. Citado 3 vezes nas páginas 30, 31, and 79.

PIRES, D. E.; ASCHER, D. B. mCSM–NA: predicting the effects of mutations on protein–nucleic acids interactions. *Nucleic Acids Research*, Oxford University Press (OUP), v. 45, n. W1, p. W241–W246, abr. 2017. Disponível em: <https://doi.org/10.1093/nar/gkx236>. Citado 4 vezes nas páginas 30, 31, 38, and 79.

PIRES, D. E. V.; ASCHER, D. B.; BLUNDELL, T. L. DUET: a server for predicting effects of mutations on protein stability using an integrated computational approach. *Nucleic Acids Research*, Oxford University Press (OUP), v. 42, n. W1, p. W314–W319, maio 2014. Disponível em: <https://doi.org/10.1093/nar/gku411>. Citado 4 vezes nas páginas 30, 31, 38, and 79.

PIRES, D. E. V.; BLUNDELL, T. L.; ASCHER, D. B. pkCSM: Predicting small-molecule pharmacokinetic and toxicity properties using graph-based signatures. *Journal of Medicinal Chemistry*, American Chemical Society (ACS), v. 58, n. 9, p. 4066–4072, abr. 2015. Disponível em: <https://doi.org/10.1021/acs.jmedchem.5b00104>. Citado 4 vezes nas páginas 30, 31, 38, and 79.

PIRES, D. E. V. et al. aCSM: noise-free graph-based signatures to large-scale receptor-based ligand prediction. *Bioinformatics*, Oxford University Press (OUP), v. 29, n. 7, p. 855–861, fev. 2013. Disponível em: <https://doi.org/10.1093/bioinformatics/btt058>. Citado 4 vezes nas páginas 30, 31, 38, and 79.

QUON, T. et al. Therapeutic opportunities and challenges in targeting the orphan g protein-coupled receptor GPR35. *ACS Pharmacology & Translational Science*, American Chemical Society (ACS), v. 3, n. 5, p. 801–812, jul. 2020. Disponível em: <https://doi.org/10.1021/acsptsci.0c00079>. Citado na página 47.

RANKOVIC, Z.; BRUST, T. F.; BOHN, L. M. Biased agonism: An emerging paradigm in GPCR drug discovery. *Bioorganic & Medicinal Chemistry Letters*, Elsevier BV, v. 26, n. 2, p. 241–250, jan. 2016. Disponível em: <https://doi.org/10.1016/j.bmcl.2015.12.024>. Citado na página 22.

RASTELLI, G.; PINZI, L. Computational polypharmacology comes of age. *Frontiers in Pharmacology*, Frontiers Media SA, v. 6, jul. 2015. Disponível em: <https://doi.org/10.3389/fphar.2015.00157>. Citado na página 79.

RATAJ, K. et al. Fingerprint-based machine learning approach to identify potent and selective 5-HT2br ligands. *Molecules*, MDPI AG, v. 23, n. 5, p. 1137, maio 2018. Disponível em: <https://doi.org/10.3390/molecules23051137>. Citado 2 vezes nas páginas 29 and 30.

ROSENBAUM, D. M.; RASMUSSEN, S. G. F.; KOBILKA, B. K. The structure and function of g-protein-coupled receptors. *Nature*, Springer Science and Business Media LLC, v. 459, n. 7245, p. 356–363, maio 2009. Disponível em: <https://doi.org/10.1038/nature08144>. Citado na página 28.

RUAT, M. et al. Targeting of smoothened for therapeutic gain. *Trends in Pharmacological Sciences*, Elsevier BV, v. 35, n. 5, p. 237–246, maio 2014. Disponível em: <https://doi.org/10.1016/j.tips.2014.03.002>. Citado na página 19.

SAKAI, M. et al. Prediction of pharmacological activities from chemical structures with graph convolutional neural networks. *Scientific Reports*, Springer Science and Business Media LLC, v. 11, n. 1, jan. 2021. Disponível em: <https://doi.org/10.1038/s41598-020-80113-7>. Citado na página 30.

SALON, J. A.; LODOWSKI, D. T.; PALCZEWSKI, K. The significance of g protein-coupled receptor crystallography for drug discovery. *Pharmacological Reviews*, American Society for Pharmacology & Experimental Therapeutics (ASPET), v. 63, n. 4, p. 901–937, out. 2011. Disponível em: <https://doi.org/10.1124/pr.110.003350>. Citado na página 16.

SANDAL, M. et al. GOMoDo: A GPCRs online modeling and docking webserver. *PLoS ONE*, Public Library of Science (PLoS), v. 8, n. 9, p. e74092, set. 2013. Disponível em: <https://doi.org/10.1371/journal.pone.0074092>. Citado na página 21.

SCHUFFENHAUER, A.; JACOBY, E. Annotating and mining the ligand-target chemogenomics knowledge space. *Drug Discovery Today: BIOSILICO*, Elsevier BV, v. 2, n. 5, p. 190–200, set. 2004. Disponível em: <https://doi.org/10.1016/s1741-8364(04)02408-4>. Citado na página 79.

SCOTT, D. J. et al. Stabilizing membrane proteins through protein engineering. *Current Opinion in Chemical Biology*, Elsevier BV, v. 17, n. 3, p. 427–435, jun. 2013. Disponível em: <https://doi.org/10.1016/j.cbpa.2013.04.002>. Citado na página 21.

SELENT, J. et al. Induced effects of sodium ions on dopaminergic g-protein coupled receptors. *PLoS Computational Biology*, Public Library of Science (PLoS), v. 6, n. 8, p. e1000884, ago. 2010. Disponível em: <https://doi.org/10.1371/journal.pcbi.1000884>. Citado na página 20.

SHIMAMURA, T. et al. Structure of the human histamine h1 receptor complex with doxepin. *Nature*, Springer Science and Business Media LLC, v. 475, n. 7354, p. 65–70, jun. 2011. Disponível em: <https://doi.org/10.1038/nature10236>. Citado na página 25.

SPASSOV, V. Z.; FLOOK, P. K.; YAN, L. LOOPER: a molecular mechanics-based algorithm for protein loop prediction. *Protein Engineering Design and Selection*, Oxford University Press (OUP), v. 21, n. 2, p. 91–100, jan. 2008. Disponível em: <https://doi.org/10.1093/protein/gzm083>. Citado na página 21.

SRIRAM, K.; INSEL, P. A. G protein-coupled receptors as targets for approved drugs: How many targets and how many drugs? *Molecular Pharmacology*, American Society for Pharmacology & Experimental Therapeutics (ASPET), v. 93, n. 4, p. 251–258, jan. 2018. Disponível em: <https://doi.org/10.1124/mol.117.111062>. Citado na página 26.

STERLING, T.; IRWIN, J. J. ZINC 15 – ligand discovery for everyone. *Journal of Chemical Information and Modeling*, American Chemical Society (ACS), v. 55, n. 11, p. 2324–2337, nov. 2015. Disponível em: <https://doi.org/10.1021/acs.jcim.5b00559>. Citado na página 29.

STRACHAN, R. T. et al. Divergent transducer-specific molecular efficacies generate biased agonism at a g protein-coupled receptor (GPCR). *Journal of Biological Chemistry*, Elsevier BV, v. 289, n. 20, p. 14211–14224, maio 2014. Disponível em: <https://doi.org/10.1074/jbc.m114.548131>. Citado na página 28.

STRADER, C. D. et al. Conserved aspartic acid residues 79 and 113 of the beta-adrenergic receptor have different roles in receptor function. *Journal of Biological Chemistry*, Elsevier BV, v. 263, n. 21, p. 10267–10271, jul. 1988. Disponível em: <https://doi.org/10.1016/s0021-9258(19)81509-0>. Citado 2 vezes nas páginas 50 and 81.

THAL, D. M. et al. Recent advances in the determination of g protein-coupled receptor structures. *Current Opinion in Structural Biology*, Elsevier BV, v. 51, p. 28–34, ago. 2018. Disponível em: <https://doi.org/10.1016/j.sbi.2018.03.002>. Citado na página 21.

TODESCHINI, R.; CONSONNI, V. *Molecular Descriptors for Chemoinformatics*. Wiley, 2009. Disponível em: <https://doi.org/10.1002/9783527628766>. Citado na página 38.

VASS, M. et al. Molecular interaction fingerprint approaches for GPCR drug discovery. *Current Opinion in Pharmacology*, Elsevier BV, v. 30, p. 59–68, out. 2016. Disponível em: <https://doi.org/10.1016/j.coph.2016.07.007>. Citado na página 81.

VIOLIN, J. D. et al. Biased ligands at g-protein-coupled receptors: promise and progress. *Trends in Pharmacological Sciences*, Elsevier BV, v. 35, n. 7, p. 308–316, jul. 2014. Disponível em: <https://doi.org/10.1016/j.tips.2014.04.007>. Citado na página 28.

WACKER, D.; STEVENS, R. C.; ROTH, B. L. How ligands illuminate GPCR molecular pharmacology. *Cell*, Elsevier BV, v. 170, n. 3, p. 414–427, jul. 2017. Disponível em: <https://doi.org/10.1016/j.cell.2017.07.009>. Citado na página 26.

WANG, C. et al. Structural basis for smoothened receptor modulation and chemoresistance to anticancer drugs. *Nature Communications*, Springer Science and Business Media LLC, v. 5, n. 1, jul. 2014. Disponível em: <https://doi.org/10.1038/ncomms5355>. Citado na página 25.

WANG, C. et al. Structure of the human smoothened receptor bound to an antitumour agent. *Nature*, Springer Science and Business Media LLC, v. 497, n. 7449, p. 338–343, maio 2013. Disponível em: <https://doi.org/10.1038/nature12167>. Citado na página 25.

WARING, M. J. et al. An analysis of the attrition of drug candidates from four major pharmaceutical companies. *Nature Reviews Drug Discovery*, Springer Science and Business Media LLC, v. 14, n. 7, p. 475–486, jun. 2015. Disponível em: <https://doi.org/10.1038/nrd4609>. Citado 2 vezes nas páginas 30 and 81.

WEBB, B.; SALI, A. Comparative protein structure modeling using MODELLER. *Current Protocols in Bioinformatics*, Wiley, v. 54, n. 1, jun. 2016. Disponível em: <https://doi.org/10.1002/cpbi.3>. Citado na página 21.

WEININGER, D. SMILES, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of Chemical Information and Modeling*, American Chemical Society (ACS), v. 28, n. 1, p. 31–36, fev. 1988. Disponível em: <https://doi.org/10.1021/ci00057a005>. Citado na página 34.

WHITE, K. L. et al. Structural connection between activation microswitch and allosteric sodium site in GPCR signaling. *Structure*, Elsevier BV, v. 26, n. 2, p. 259–269.e5, fev. 2018. Disponível em: <https://doi.org/10.1016/j.str.2017.12.013>. Citado na página 20.

WISHART, D. S. et al. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Research*, Oxford University Press (OUP), v. 46, n. D1, p. D1074–D1082, nov. 2017. Disponível em: <https://doi.org/10.1093/nar/gkx1037>. Citado na página 29.

WON, J. et al. GalaxyGPCRloop: Template-based and ab initio structure sampling of the extracellular loops of g-protein-coupled receptors. *Journal of Chemical Information and Modeling*, American Chemical Society (ACS), v. 58, n. 6, p. 1234–1243, maio 2018. Disponível em: <https://doi.org/10.1021/acs.jcim.8b00148>. Citado na página 21.

WOOLLEY, M. J.; CONNER, A. C. Understanding the common themes and diverse roles of the second extracellular loop (ECL2) of the GPCR super-family. *Molecular and Cellular Endocrinology*, Elsevier BV, v. 449, p. 3–11, jul. 2017. Disponível em: <https://doi.org/10.1016/j.mce.2016.11.023>. Citado na página 21.

WORTH, C. L. et al. GPCR-SSFE: A comprehensive database of g-protein-coupled receptor template predictions and homology models. *BMC Bioinformatics*, Springer Science and Business Media LLC, v. 12, n. 1, maio 2011. Disponível em: <https://doi.org/10.1186/1471-2105-12-185>. Citado na página 21.

WU, B. et al. Structures of the CXCR4 chemokine GPCR with small-molecule and cyclic peptide antagonists. *Science*, American Association for the Advancement of Science (AAAS), v. 330, n. 6007, p. 1066–1071, nov. 2010. Disponível em: <https://doi.org/10.1126/science.1194396>. Citado na página 25.

WU, H. et al. Structure of a class c GPCR metabotropic glutamate receptor 1 bound to an allosteric modulator. *Science*, American Association for the Advancement of Science (AAAS), v. 344, n. 6179, p. 58–64, abr. 2014. Disponível em: <https://doi.org/10.1126/science.1249489>. Citado na página 25.

WU, J. et al. Precise modelling and interpretation of bioactivities of ligands targeting g protein-coupled receptors. *Bioinformatics*, Oxford University Press (OUP), v. 35, n. 14, p. i324–i332, jul. 2019. Disponível em: <https://doi.org/10.1093/bioinformatics/btz336>. Citado 4 vezes nas páginas 30, 34, 64, and 80.

WU, J. et al. WDL-RF: predicting bioactivities of ligand molecules acting with g protein-coupled receptors by combining weighted deep learning and random forest. *Bioinformatics*, Oxford University Press (OUP), v. 34, n. 13, p. 2271–2282, fev. 2018. Disponível em: <https://doi.org/10.1093/bioinformatics/bty070>. Citado 12 vezes nas páginas 9, 30, 34, 44, 45, 46, 64, 66, 67, 70, 71, and 80.

YAMAGUCHI, S.; KANEKO, M.; NARUKAWA, M. Approval success rates of drug candidates based on target, action, modality, application, and their combinations. *Clinical and Translational Science*, Wiley, v. 14, n. 3, p. 1113–1122, abr. 2021. Disponível em: <https://doi.org/10.1111/cts.12980>. Citado na página 30.

YANG, D. et al. G protein-coupled receptors: structure- and function-based drug discovery. *Signal Transduct. Target. Ther.*, Springer Science and Business Media LLC, v. 6, n. 1, p. 7, jan. 2021. Citado na página 17.

YANG, D. et al. G protein-coupled receptors: structure- and function-based drug discovery. *Signal Transduction and Targeted Therapy*, Springer Science and Business Media LLC, v. 6, n. 1, jan. 2021. Disponível em: <https://doi.org/10.1038/s41392-020-00435-w>. Citado na página 20.

ZHANG, J. et al. GPCR-i-TASSER: A hybrid approach to g protein-coupled receptor structure modeling and the application to the human genome. *Structure*, Elsevier BV, v. 23, n. 8, p. 1538–1549, ago. 2015. Disponível em: <https://doi.org/10.1016/j.str.2015.06.007>. Citado 5 vezes nas páginas 21, 24, 26, 46, and 57.

ZHANG, R.; XIE, X. Tools for GPCR drug discovery. *Acta Pharmacologica Sinica*, Springer Science and Business Media LLC, v. 33, n. 3, p. 372–384, jan. 2012. Disponível em: <https://doi.org/10.1038/aps.2011.173>. Citado 2 vezes nas páginas 16 and 25.

ZHOU, Q. et al. Common activation mechanism of class a GPCRs. *eLife*, eLife Sciences Publications, Ltd, v. 8, dez. 2019. Disponível em: <https://doi.org/10.7554/elife.50279>. Citado na página 20.

ZHU, S. et al. Trends in application of advancing computational approaches in GPCR ligand discovery. *Experimental Biology and Medicine*, SAGE Publications, v. 246, n. 9, p. 1011–1024, fev. 2021. Disponível em: <https://doi.org/10.1177/1535370221993422>. Citado na página 28.

ZIN, P. P. K.; WILLIAMS, G. J.; EKINS, S. Cheminformatics analysis and modeling with MacrolactoneDB. *Scientific Reports*, Springer Science and Business Media LLC, v. 10, n. 1, abr. 2020. Disponível em: <https://doi.org/10.1038/s41598-020-63192-4>. Citado 3 vezes nas páginas 34, 64, and 80.

Appendix

Table A.1: Programming and Scripting tools used in this work.

Task	**Tool**	**Description**
Scripting	Python	An interpreted high-level programming used for general purposes.
Data Manipulation	Numpy	Is a library for Python which provides support for large, multi-dimensional arrays, matrices, and high-level mathematical functions.
Data Analysis	Pandas	It is a library for Python which allows data manipulation and analysis, such as merging, reshaping, selecting, data cleaning and many others.
Data Analysis	Scikit-Learn	It is a free software machine learning library for Python. It features various classification, regression and clustering algorithms.
Web Framework	Flask	Python framework used for the development of web servers.
Web Development	HTML	HyperText Markup Language is the standard markup language for creating websites.
Web Development	CSS	It is a style sheet language used for describing the presentation of a document written in HTML.
Web Development	JavaScript	JavaScript is used for controlling the behaviour of elements on an HTML page.
Containers	Anaconda	Anaconda is a distribution of Python and R that aims to simplify package management and deployment.

Script A.1: Script used for data set curation.

```
import pandas as pd
import numpy as np

#Import data set
ds_df = pd.read_csv("Data set to be curated file")

#Drop entries which smiles could not be recovered
ds_df_smilesFull=ds_df.dropna(subset=['smiles'])

#Sort entries according to CID
ds_df_sorted = ds_df_smilesFull.sort_values(by=["cid"])

#Select only entries with activity information:
ds_df_class_available_pre= ds_df_sorted[ds_df_sorted['activity'].str.match('Active|Inactive')]

#Take out dubious entries:
#Select entries with Active and Inactive classification and keep only one entry for each:
ds_controversy=(ds_df_class_available_pre.groupby(['cid'])['activity'].nunique()==2)
list_of_contoversy = ds_controversy[ds_controversy].index.values
ds_df_class_available =
ds_df_class_available_pre[~ds_df_class_available_pre.cid.isin(list_of_contoversy)]

#Prepare data set for regression:
#Delete entries without activity values:
ds_df_acvalueFull= ds_df_class_available.dropna(subset=['acvalue'])

#Select entries with type of activity = Ki, Kd, IC50 or EC50:
ds_df_regre_pre= ds_df_acvalueFull[ds_df_acvalueFull['acname'].str.match('Ki|Kd|IC50|EC50',
na=False)]

#Select repeated entries and keep just one:
ds_df_acvalueFull_sorted = ds_df_regre_pre.sort_values(['cid', 'acname'], ascending=[True,
False])
ds_df_regression_without_repeats = ds_df_acvalueFull_sorted.drop_duplicates(subset='cid',
keep='first')
ds_df_regression_without_repeats_all=
ds_df_regression_without_repeats[ds_df_regression_without_repeats['acname'].str.match('IC50|
EC50|Ki|Kd')][['smiles','acvalue']]

#Convert activity values for -log[Molar]:
ds_df_regression_without_repeats_all['minuslog_acvalue'] =
(-1*np.log10(ds_df_regression_without_repeats_all.acvalue))
del ds_df_regression_without_repeats_all['acvalue']

#Save final dataset
ds_df_regression_without_repeats_all.to_csv('Output data set file'', index=False, sep='\t')
```

Script A.2: Script used for generating blind test

```
#import modules
from rdkit import Chem
from rdkit.Chem import AllChem
from rdkit import DataStructs
import pandas as pd
from rdkit.ML.Cluster import Butina
from random import sample
import itertools
import seaborn as sns
from scipy.stats import ks_2samp
import matplotlib.pyplot as plt
#import numpy as np
import sys

inputfile1 = sys.argv[1] #input file with smiles
receptor_id = sys.argv[2]

out_test_df= inputfile1+"_test_df.csv"
out_train_df= inputfile1+"_train_df.csv"
out_graph= inputfile1+"_graph.png"

def generate_dataset_for_training():
        #Function which receives fingerprints from the smiles and clusters the data:
        def ClusterFps(fps,cutoff=0.2):
        # first generate the distance matrix:
        dists = []
        nfps = len(fps)
        for i in range(1,nfps):
        #compare all smiles with all smiles and generate a smilarity matrix:
        sims = DataStructs.BulkTanimotoSimilarity(fps[i],fps[:i])
        #Convert similarity for 1-similarity
        dists.extend([1-x for x in sims])

        # now cluster the data:
        cs = Butina.ClusterData(dists,nfps,cutoff,isDistData=True)
        #print(cs)
        return cs
        ####End of the function

        #Open the input as pandas dataframe
        smiles_df = pd.read_csv(inputfile1, sep = ',')

        #Select only smiles column
        df_smiles = smiles_df['SMILES']

        #Create an empty list which will receive the smiles
        c_smiles = []
        #Appending smiles to the previous list
        for ds in df_smiles:
        c_smiles.append(ds)

        #Estimating how many smiles will be necessary to the blind test (10%)
        number_smiles_blind_test = (len(c_smiles))/10
```

```
        # make a list of mols using the list of smiles for fingerprint generation
        ms = [Chem.MolFromSmiles(x) for x in c_smiles]

        # make a list of fingerprints (fp), Morgan fingerprints
        fps = [AllChem.GetMorganFingerprintAsBitVect(m, 2) for m in ms]

        #Clustering data, the results are multiples tuples containing the index of the mols
        clusters=ClusterFps(fps,cutoff=0.25)

        #Create an empty tuple and an empty lists which will receive the mols indexd for the
blind test, at last only the list will be used
        smiles_for_blind_test_id = [] #tuple
        smiles_for_blind_test_id2 =[] #list

        #Sort the clusters in crescent order using the number of smiles inside:
        clusters = (sorted(clusters, key=len))
        #Iterates over the tuples to append some of them in order to get 10% for blind test
        for x in clusters:
        #Before appending check if the number of smiles (id) reached 10% plus a tolerance in
order to avoid infinite running
        if (len(smiles_for_blind_test_id2))<(number_smiles_blind_test+(2*len(x)+1)):

        #append clusters for the blind test
        smiles_for_blind_test_id.append(x)
        #Convert tuple for list
        smiles_for_blind_test_id2 = list(itertools.chain(*smiles_for_blind_test_id))

        #Create an empty list which will receive the smiles for the blind test
        smiles_for_blind_test_ori = []

        #Using the list of index, extract the smiles from the original list of smiles extract from the
dataset('c_smiles') and append it
        for x in smiles_for_blind_test_id2:
        smiles_for_blind_test_ori.append(c_smiles[x])

        #calculates de the number of smiles and clusters
        number_of_smiles=("Number of smiles: " +str(len(c_smiles)))
        number_of_clusters=("Number of clusters: " +str(len(clusters)))


        #Using the selceted smiles extract row from orignal data spliting the train and the test
dataset
        df_train = smiles_df[~smiles_df ['SMILES'].isin(smiles_for_blind_test_ori)]
        df_test = smiles_df[smiles_df ['SMILES'].isin(smiles_for_blind_test_ori)]

        #save both datasets
        df_train.to_csv(out_train_df, index=False, encoding='utf-8', sep = '\t')
        df_test.to_csv(out_test_df, index=False, encoding='utf-8', sep = '\t')

        #calculates the number of smiles for the test and train dataset:
        number_of_smiles4blind = ("Number for blind test: " +str(len(smiles_for_blind_test_id2)))
        number_of_smiles4test = ("Number for train test: " +str(len(df_train)))

        #Calculates Kolmogorov-Smirnov statistic and its respective p-value
```

```
        ks =
ks_2samp(df_train['minuslog_acvalue_minuslog_molar'],df_test['minuslog_acvalue_minuslog_m
olar'])
        pvalue=(ks[1])

        pvalueText= ("p-value= " + str(pvalue))

        #Generates the figures
        plt.figure(figsize=(12, 7))
        sns.distplot(df_train['minuslog_acvalue_minuslog_molar'], color='red', label='Train', kde=
True, hist = True)
        sns.distplot(df_test['minuslog_acvalue_minuslog_molar'], color='blue', label='Test', kde=
True, hist = True)
        plt.title("Datasets - Train x Test - "+ receptor_id )
        plt.xlabel('Activity- (-log)')
        plt.ylabel('Percentage of molecules')
        plt.legend()
        plt.text(3.4, 0.25, pvalueText, fontsize=10)
        plt.text(8, 0.2, (number_of_smiles) + "\n"+ (number_of_clusters)+
"\n"+number_of_smiles4test+ "\n"+number_of_smiles4blind, style='italic',
        bbox={'facecolor': 'white', 'alpha': 0.5, 'pad': 10})

        plt.savefig(out_graph)
        #plt.show()
        plt.close()
        #return pvalue
        return(pvalue)

#run the function for the first time
p_value = float(generate_dataset_for_training())

#keeps running the function until the distributions are satisfactory
while (p_value) <= (0.0):
        p_value = generate_dataset_for_training()
```

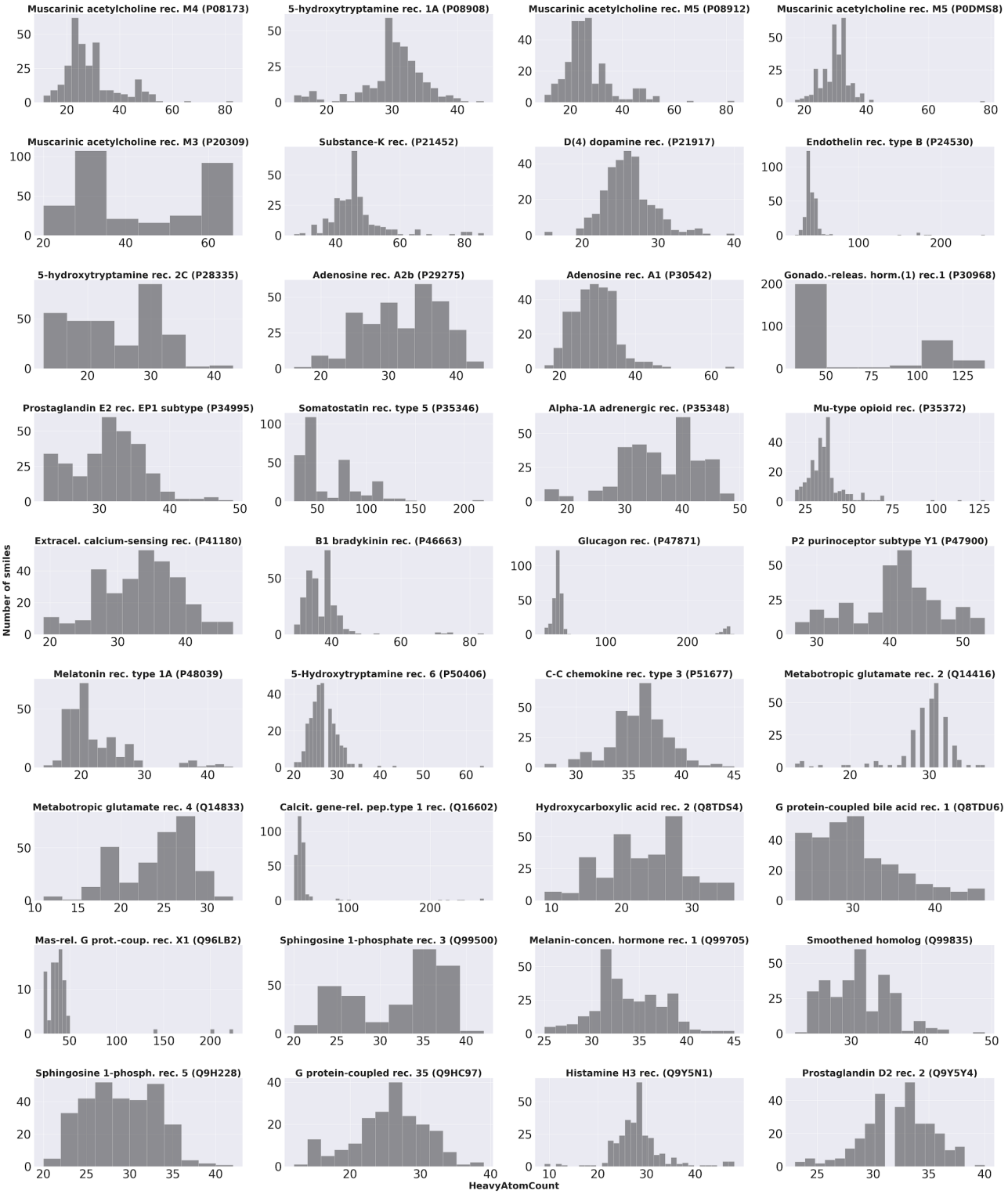


Figure A1: Potent ligands - Histograms considering heavy atoms count distribution for all datasets.

Figure A2: Potent ligands - Histograms considering number of heteroatoms distribution for all datasets.

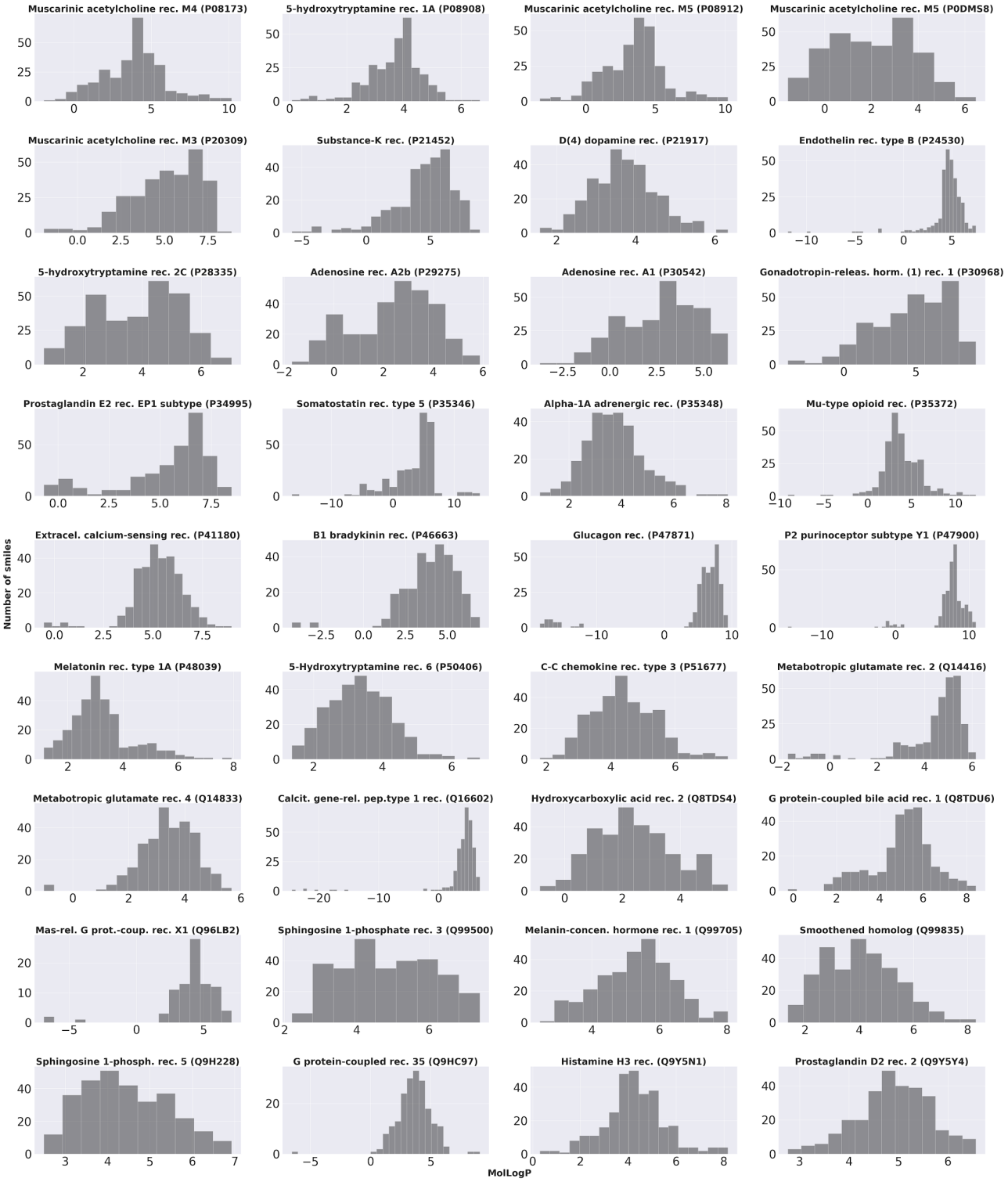


Figure A3: Potent ligands - Histograms considering log *P* distribution for all datasets.

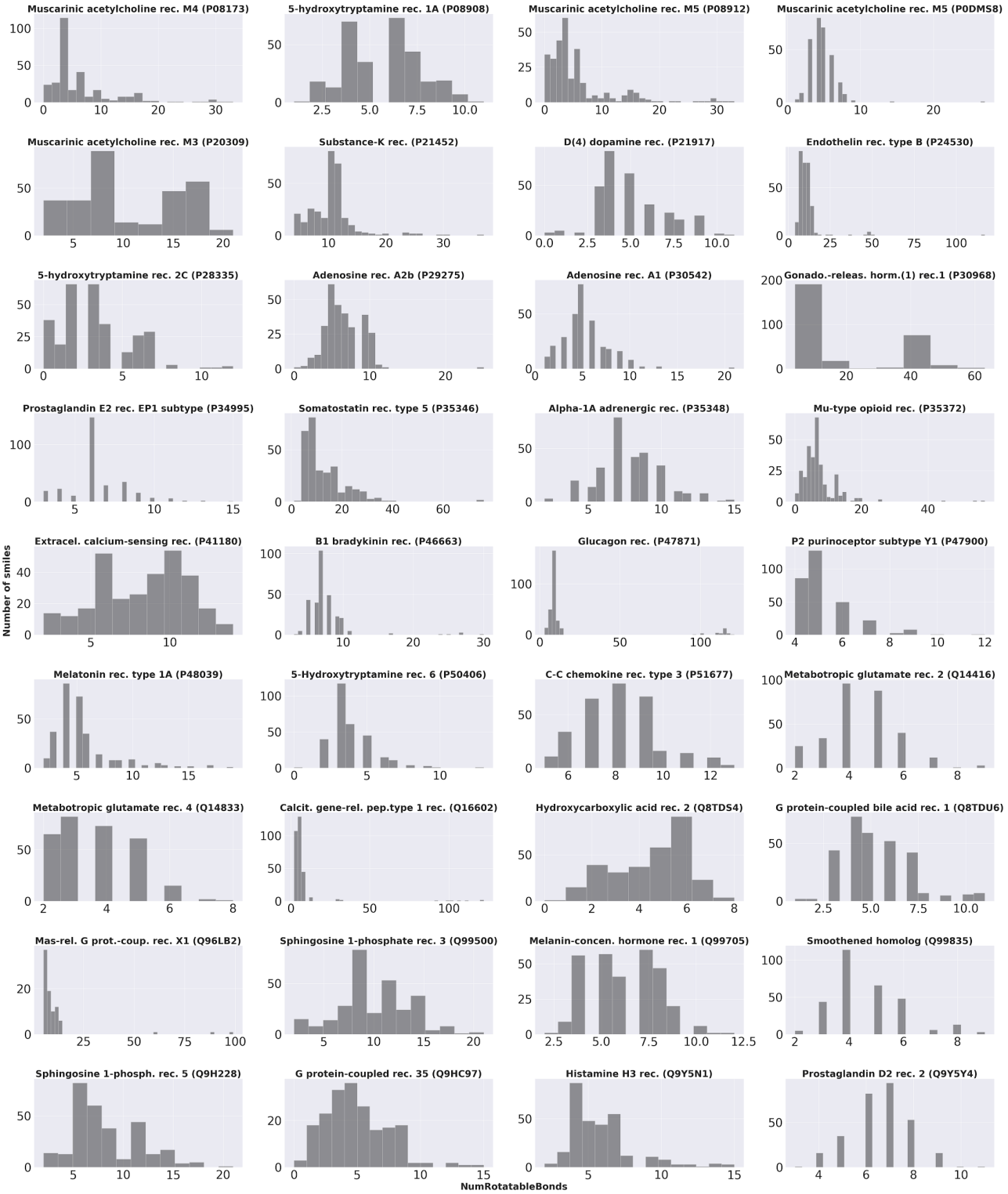


Figure A4: Potent ligands - Histograms considering number of rotatable bonds distribution for all datasets.

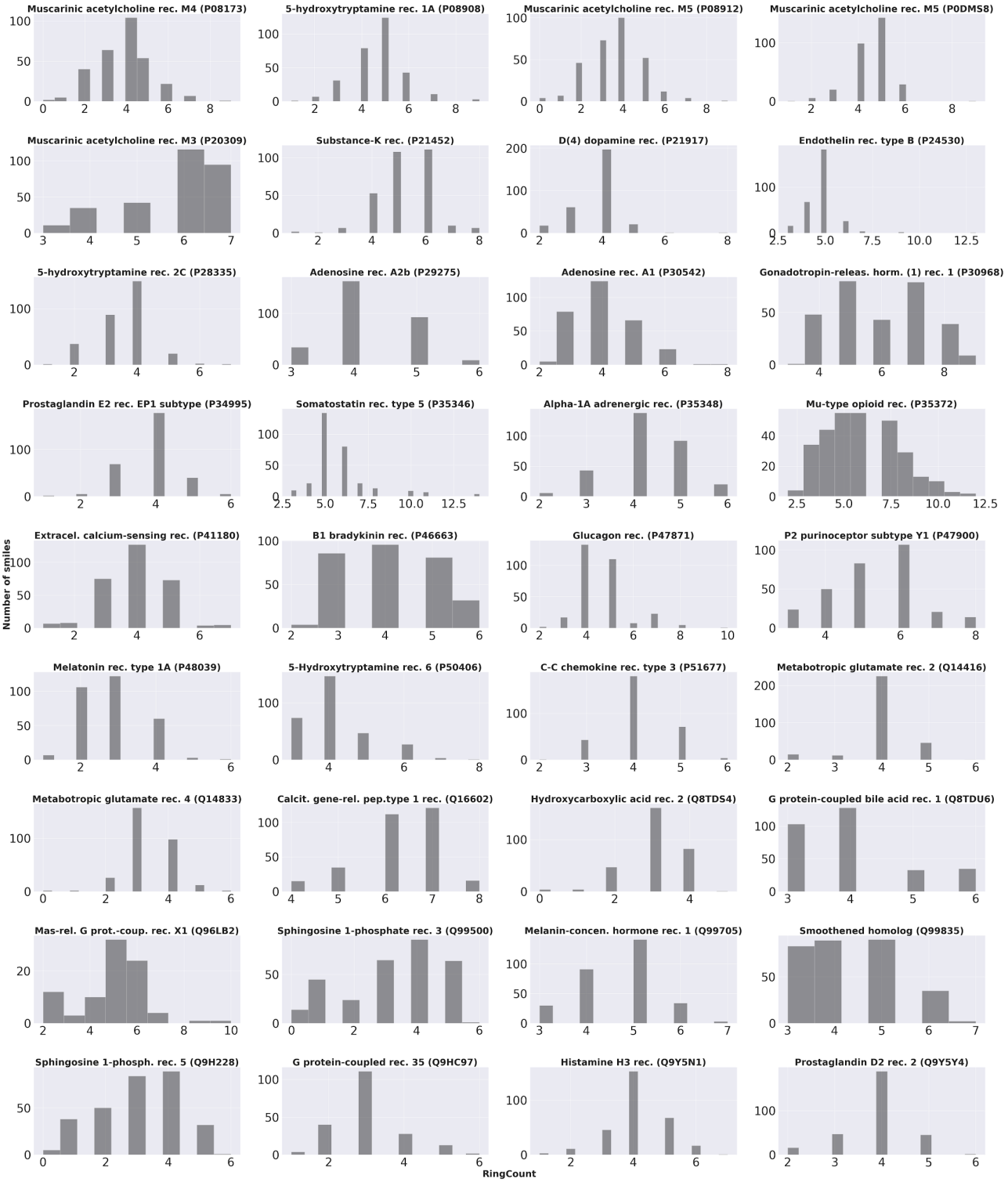


Figure A5: Potent ligands - Histograms considering number of rings count distribution for all datasets.

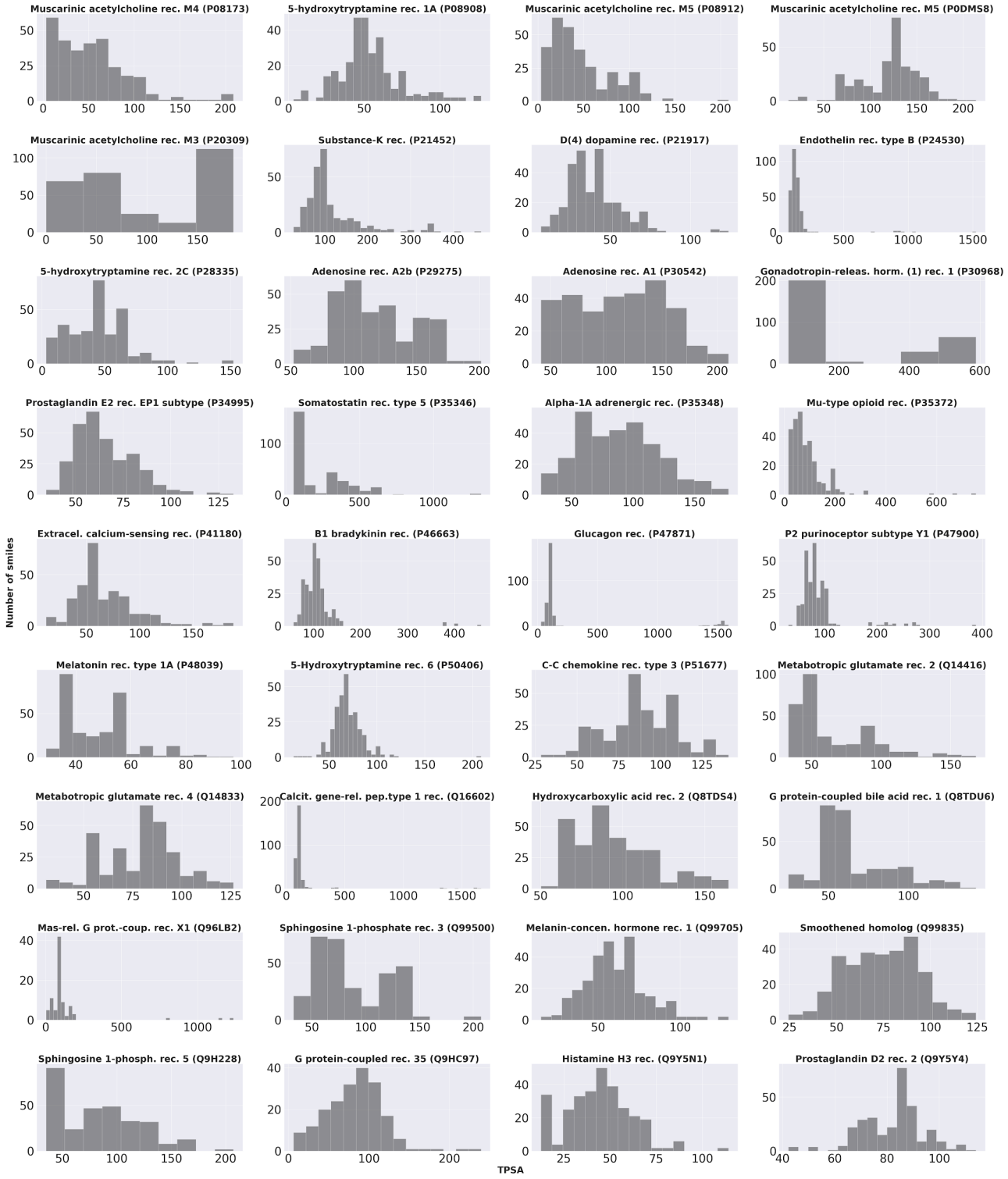


Figure A6: Potent ligands - Histograms considering topological polar surface distribution for all datasets.

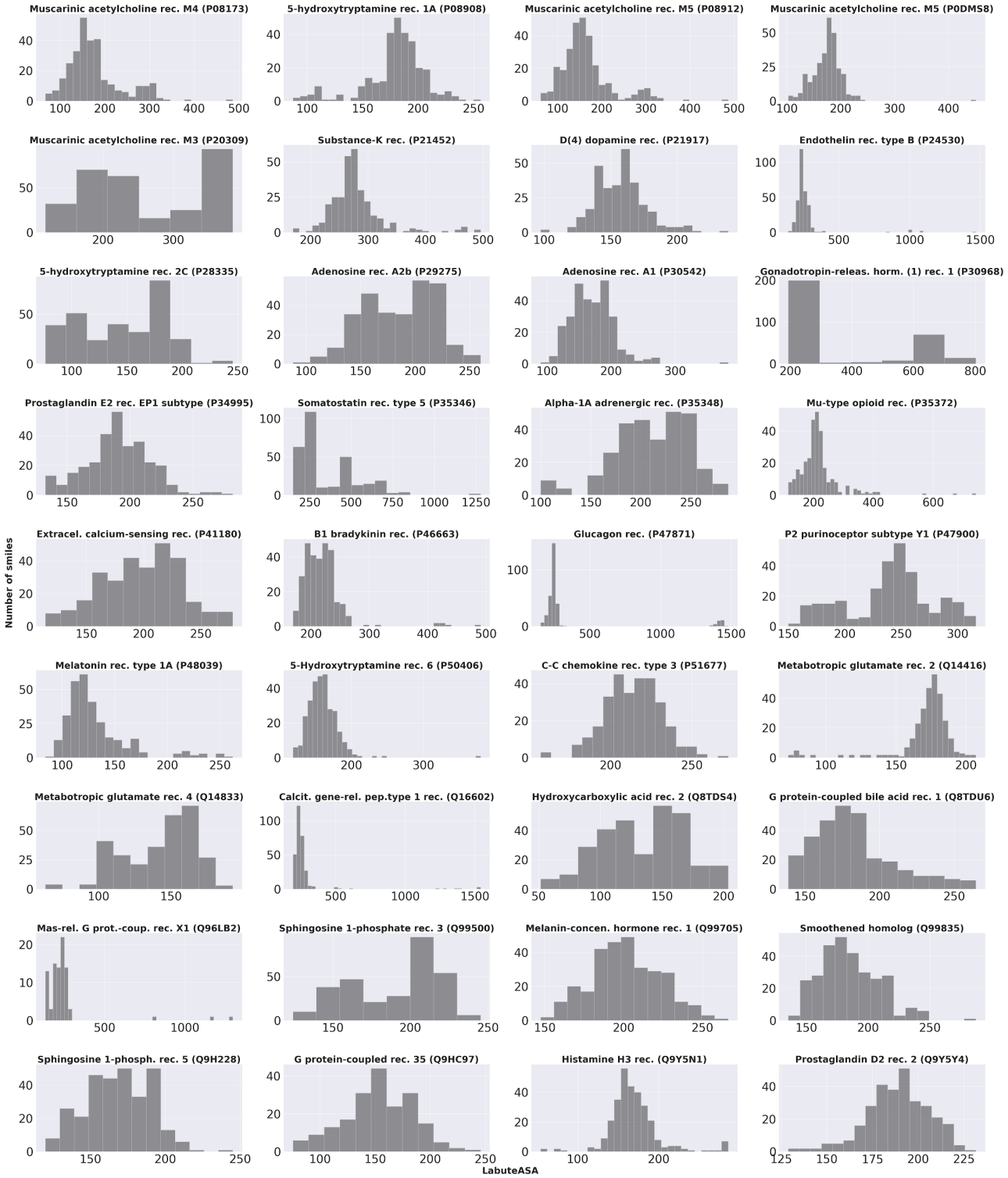


Figure A7: Potent ligands - Histograms considering Labute's Approximate Surface Area distribution for all datasets.

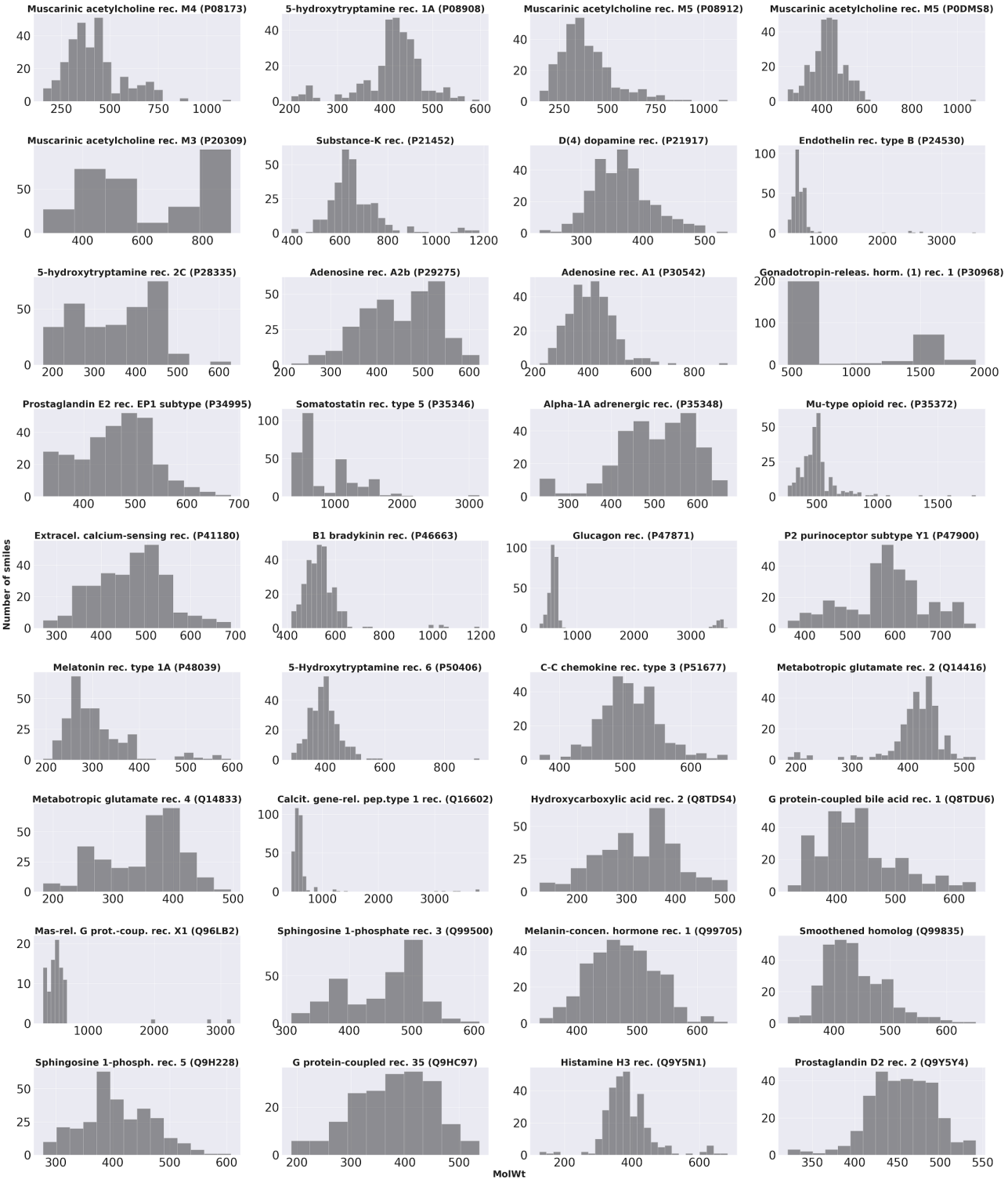


Figure A8: Potent ligands - Histograms considering molecular weight distribution for all datasets.

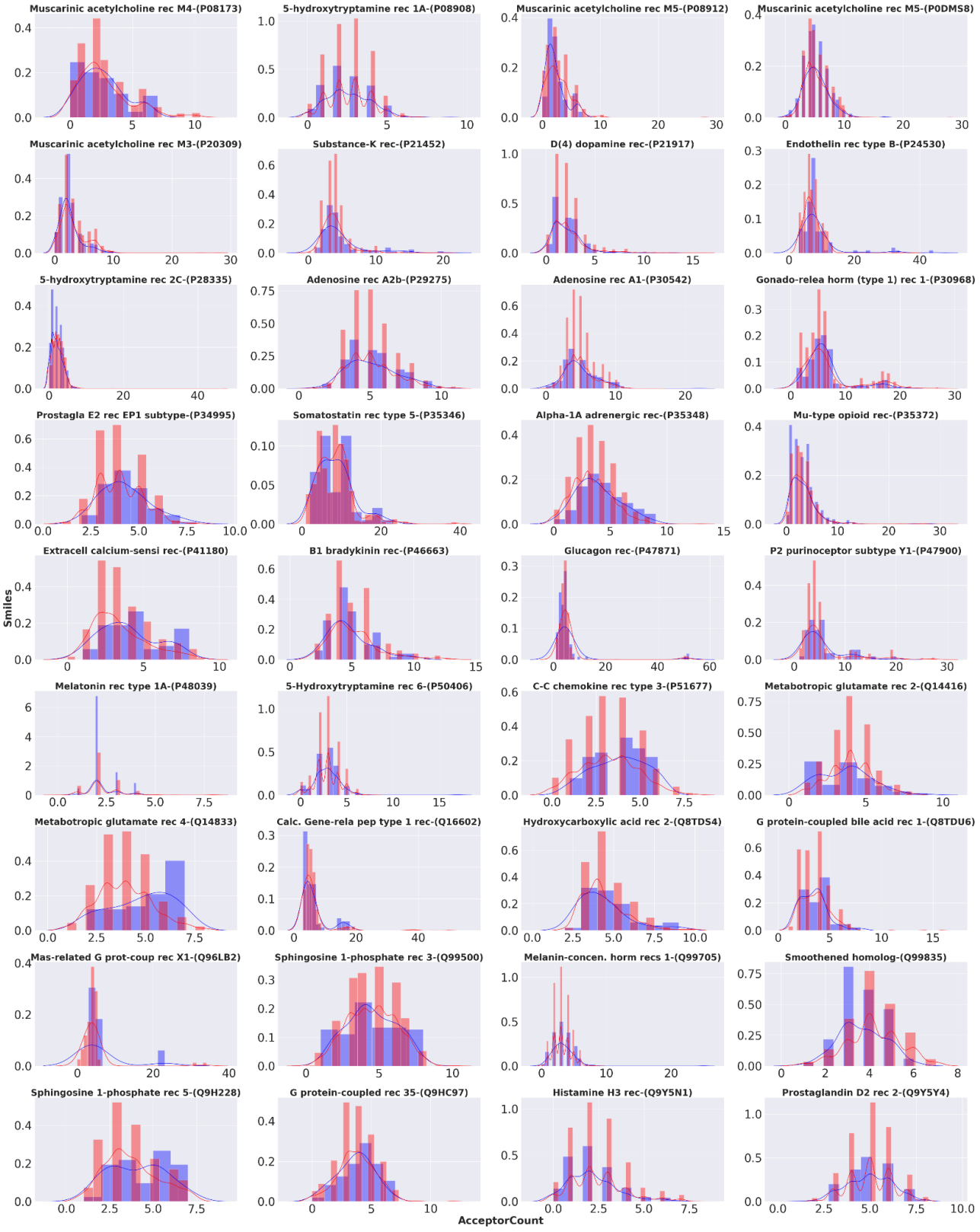


Figure A9: Histograms considering count of hydrogen bonds acceptor distribution for datasets without outliers (from cross-validation schemes) in red and considering only outliers in blue.

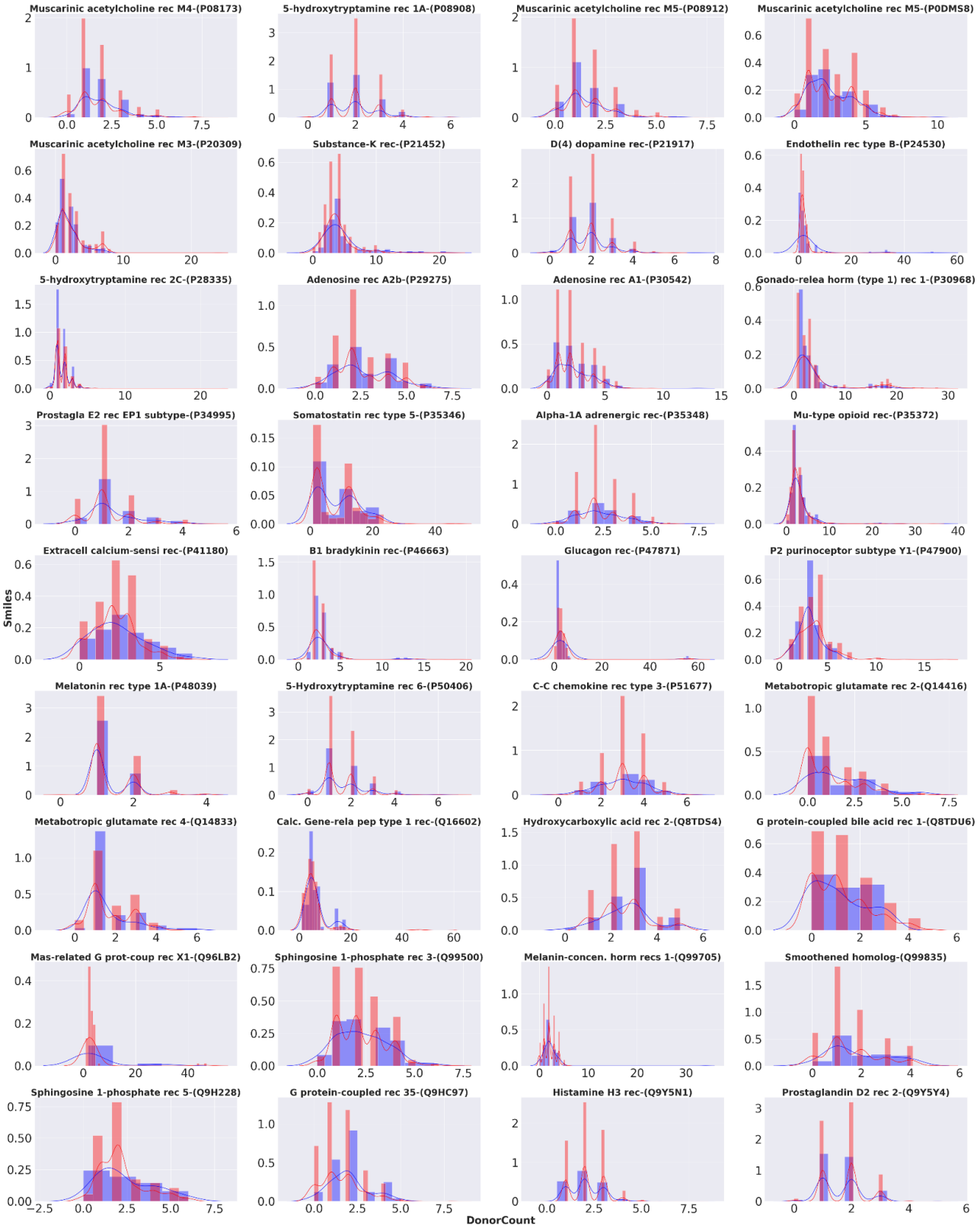


Figure A10: Histograms considering count of hydrogen bonds donor distribution for datasets without outliers (from cross-validation schemes) in red and considering only outliers in blue.

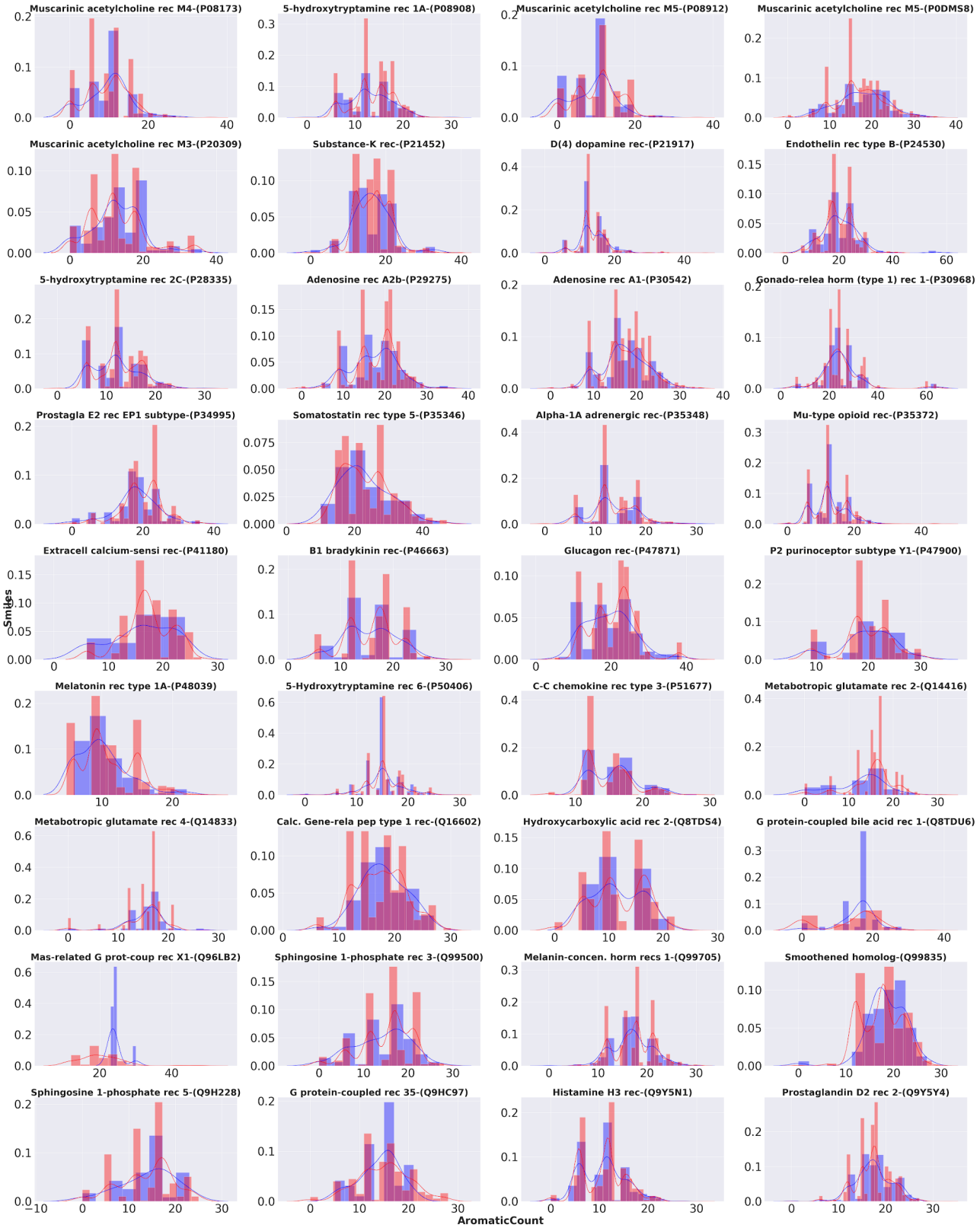


Figure A11: Histograms considering aromaticity distribution for datasets without outliers (from cross-validation schemes) in red and considering only outliers in blue.

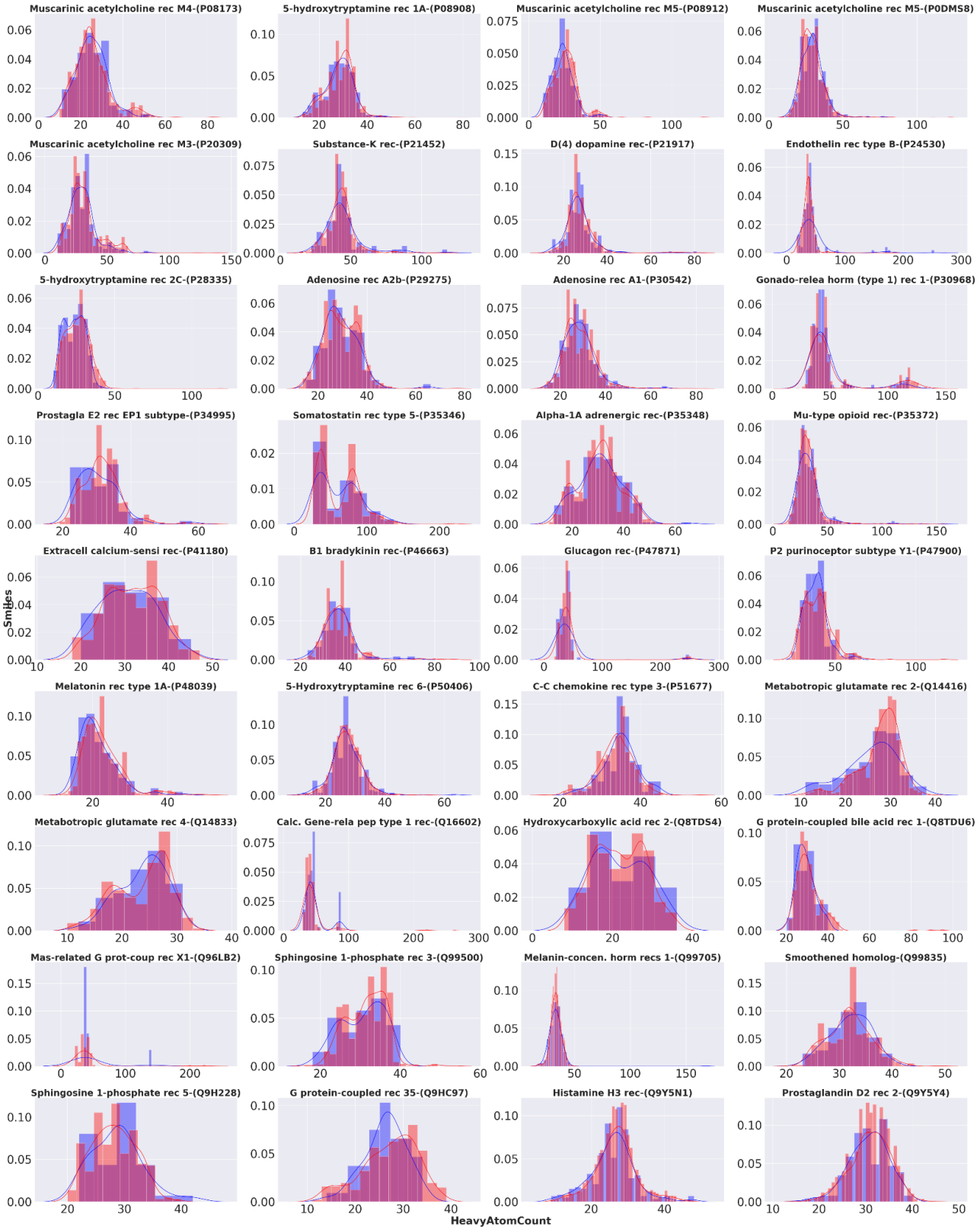


Figure A12: Histograms considering count of heavy atoms distribution for datasets without outliers (from cross-validation schemes) in red and considering only outliers in blue.

Figure A13: Histograms considering hydrophobicity distribution for datasets without outliers (from cross-validation schemes) in red and considering only outliers in blue.

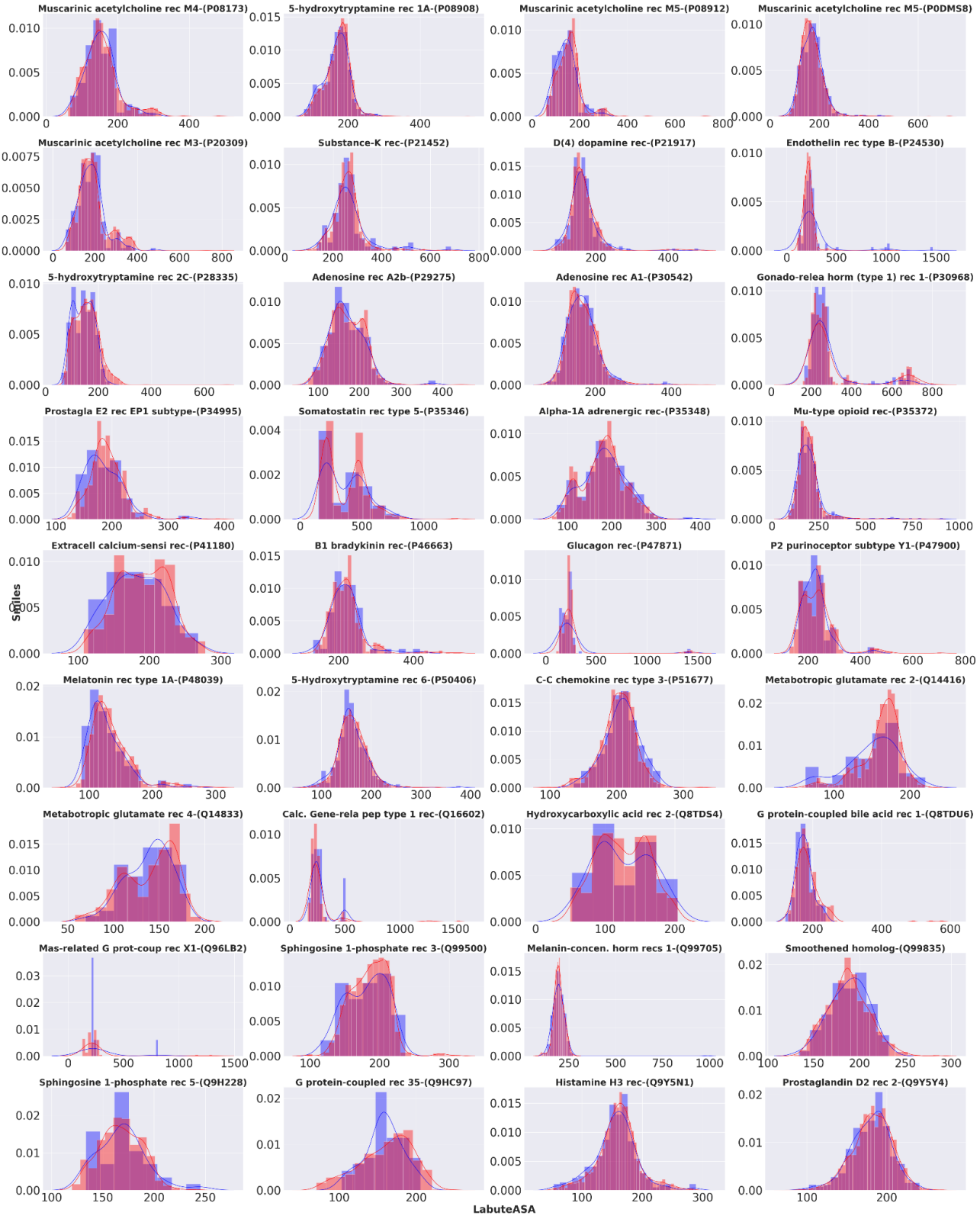


Figure A14: Histograms considering Labute's Approximate Surface Area distribution for datasets without outliers (from cross-validation schemes) in red and considering only outliers in blue.

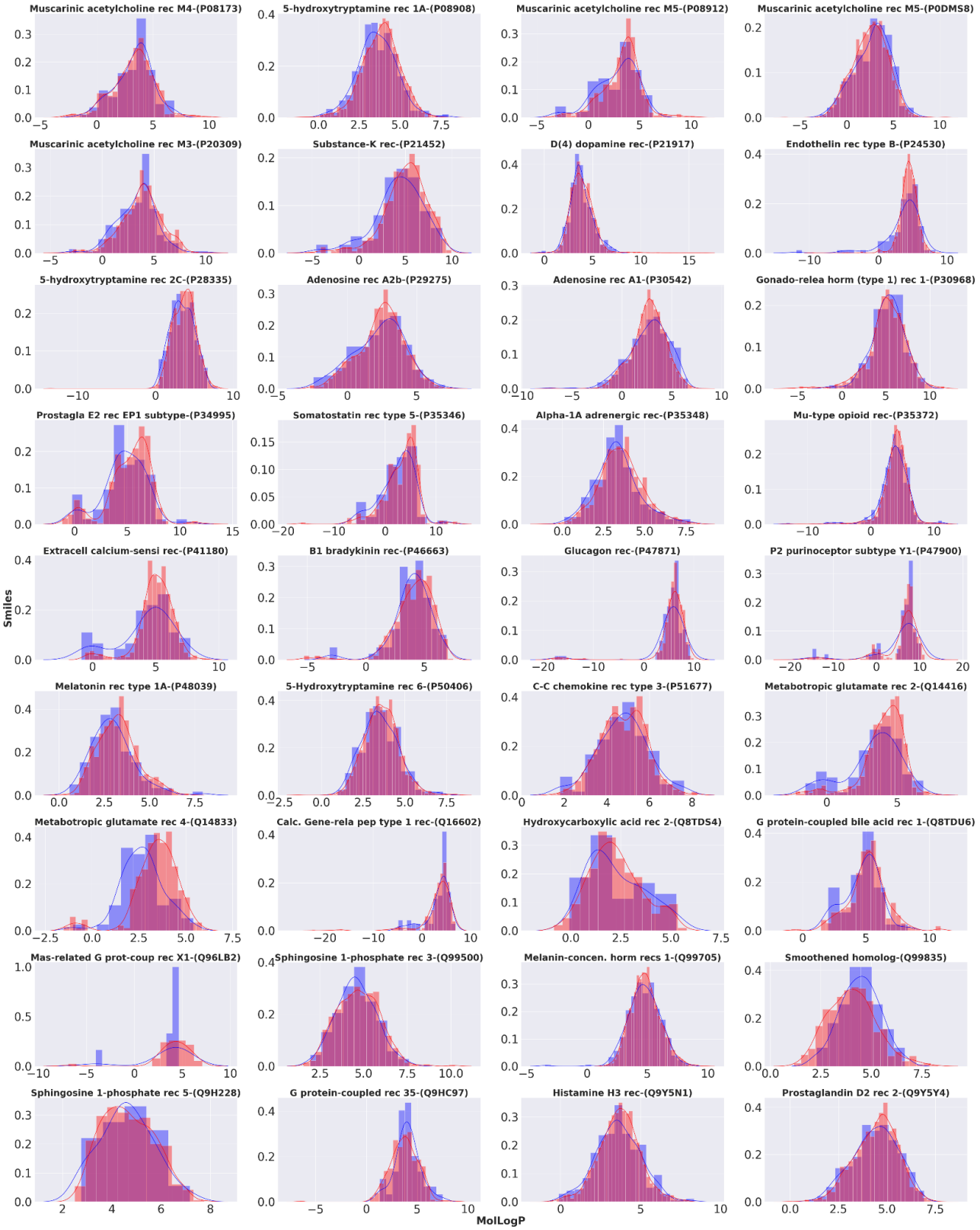


Figure A15: Histograms considering log *P* distribution for datasets without outliers (from cross-validation schemes) in red and considering only outliers in blue.

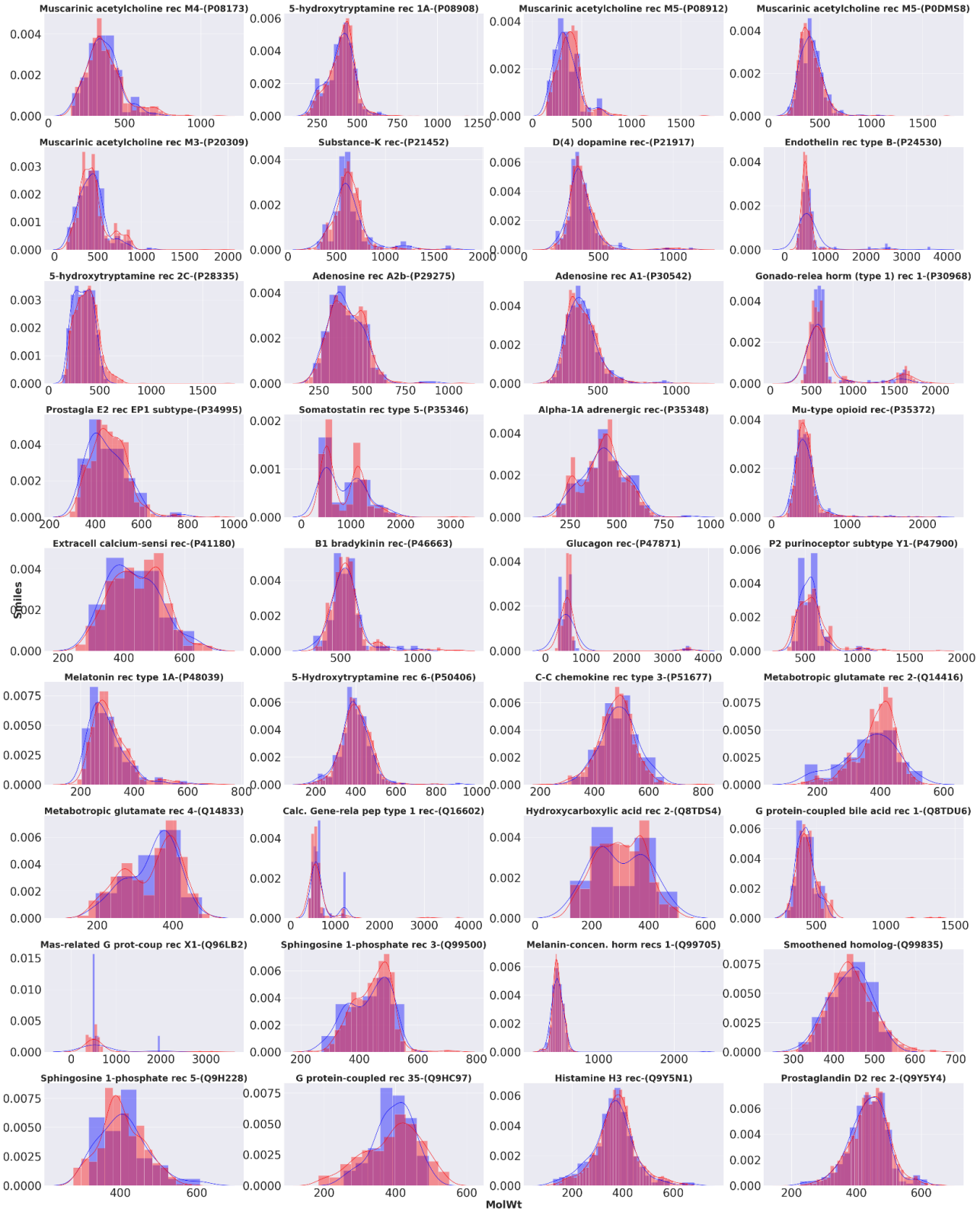


Figure A16: Histograms considering molecular weight distribution for datasets without outliers (from cross-validation schemes) in red and considering only outliers in blue.

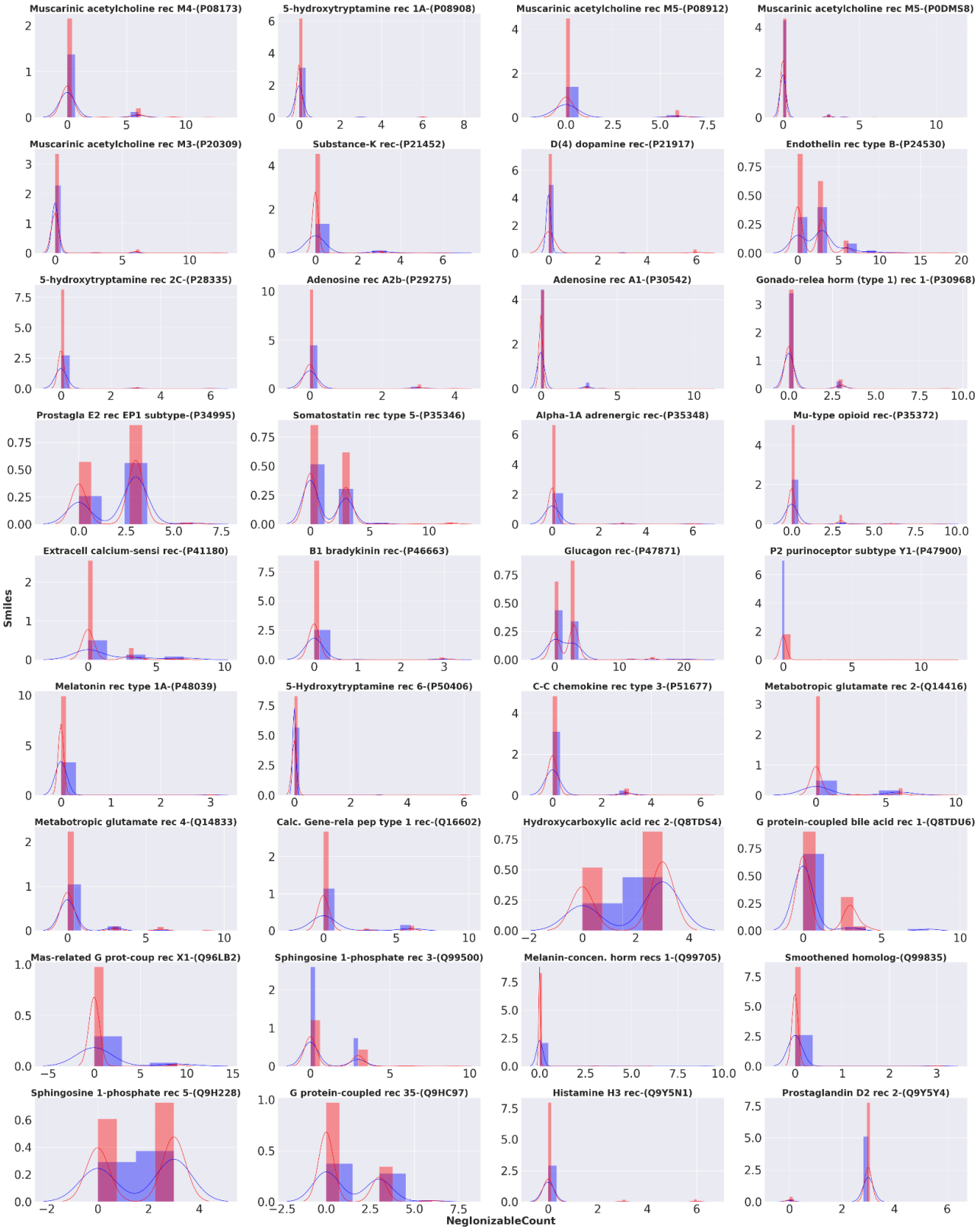


Figure A17: Histograms considering count of negative ionizable atoms for datasets without outliers (from cross-validation schemes) in red and considering only outliers in blue.

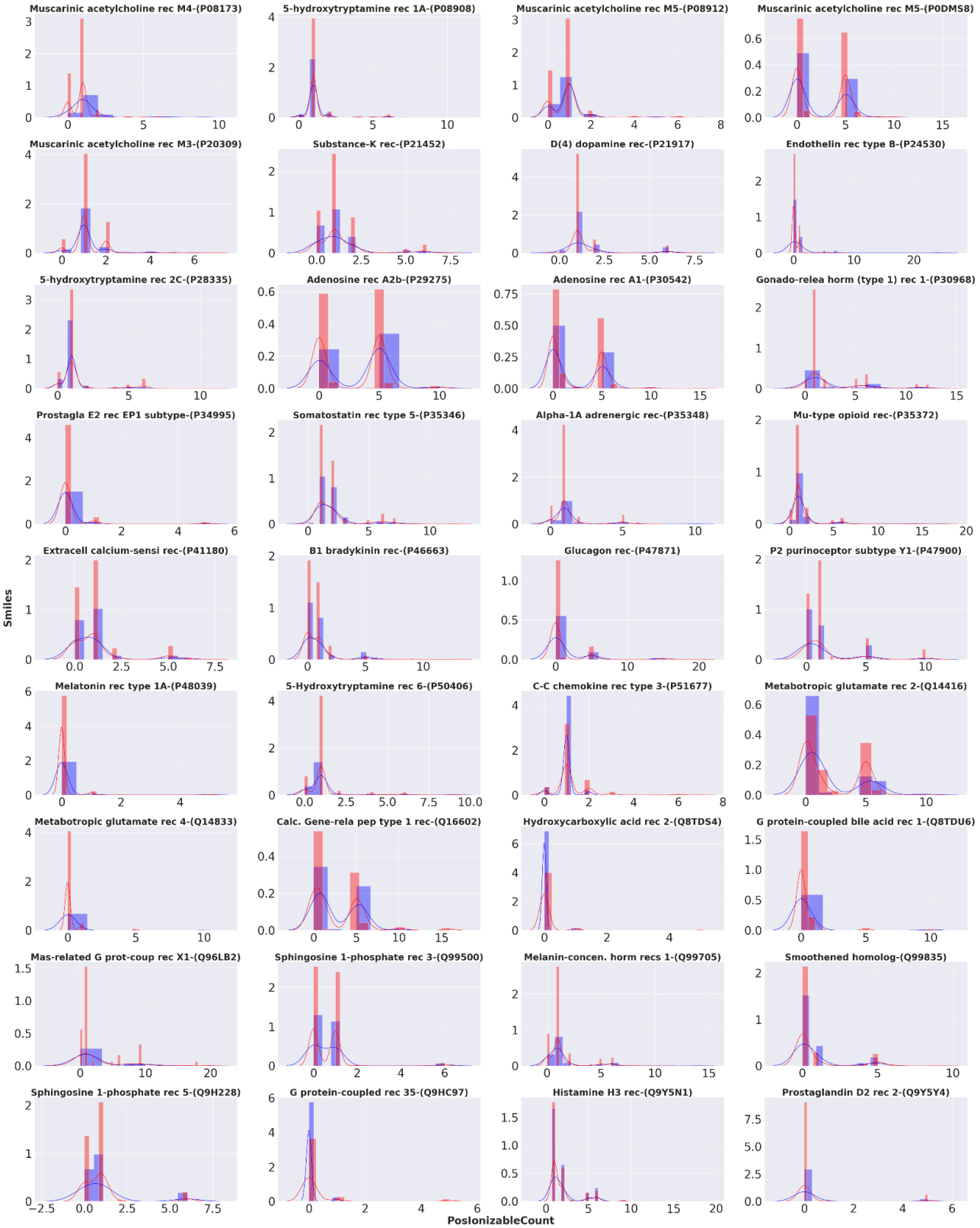


Figure A18: Histograms considering count of positive ionizable atoms for datasets without outliers (from cross-validation schemes) in red and considering only outliers in blue.

Figure A19: Histograms considering count of Rotatable bonds distribution for datasets without outliers (from cross-validation schemes) in red and considering only outliers in blue.

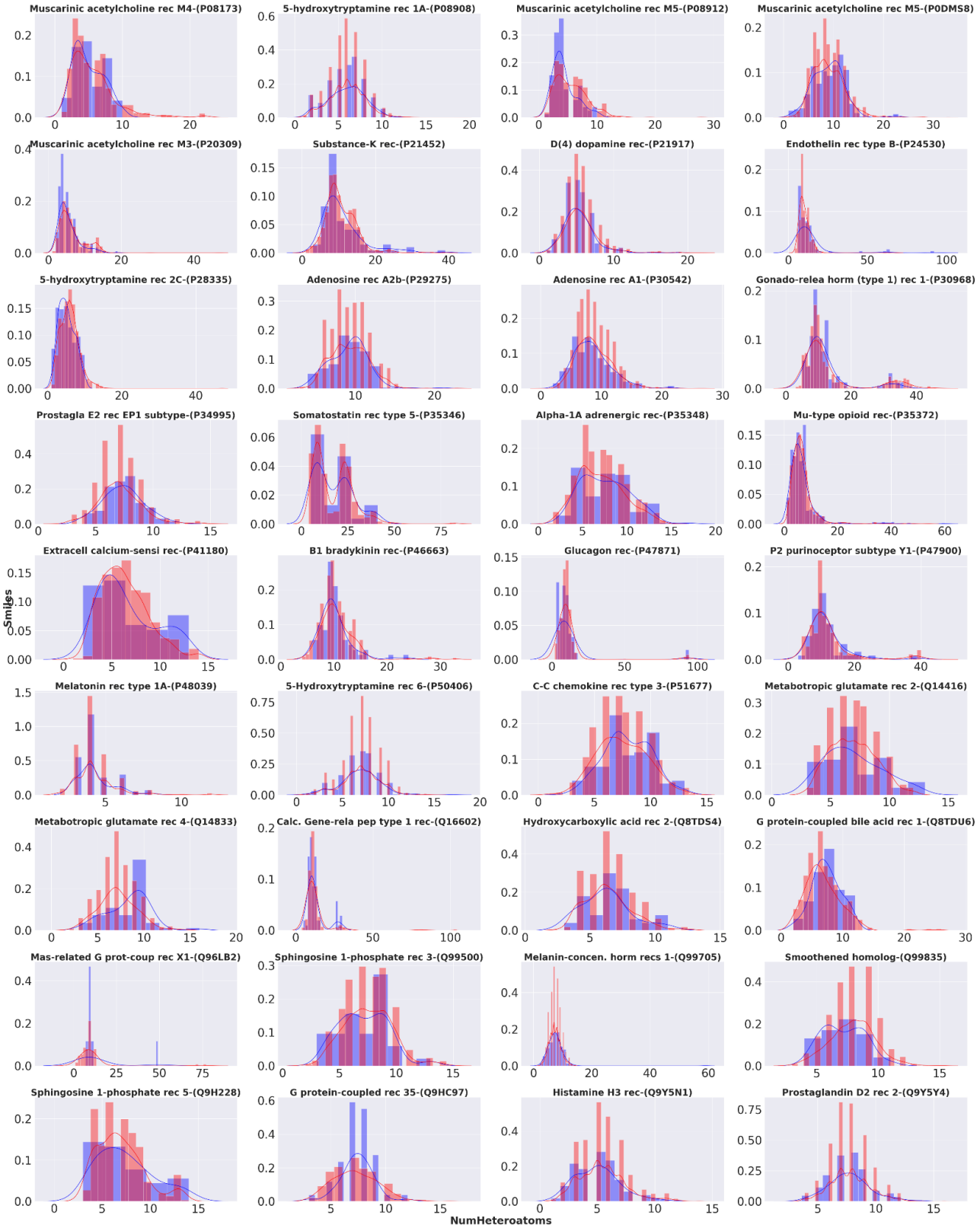


Figure A20: Histograms considering count of heteroatoms for datasets without outliers (from cross-validation schemes) in red and considering only outliers in blue.

Figure A21: Histograms considering count of rings for datasets without outliers (from cross-validation schemes) in red and considering only outliers in blue.

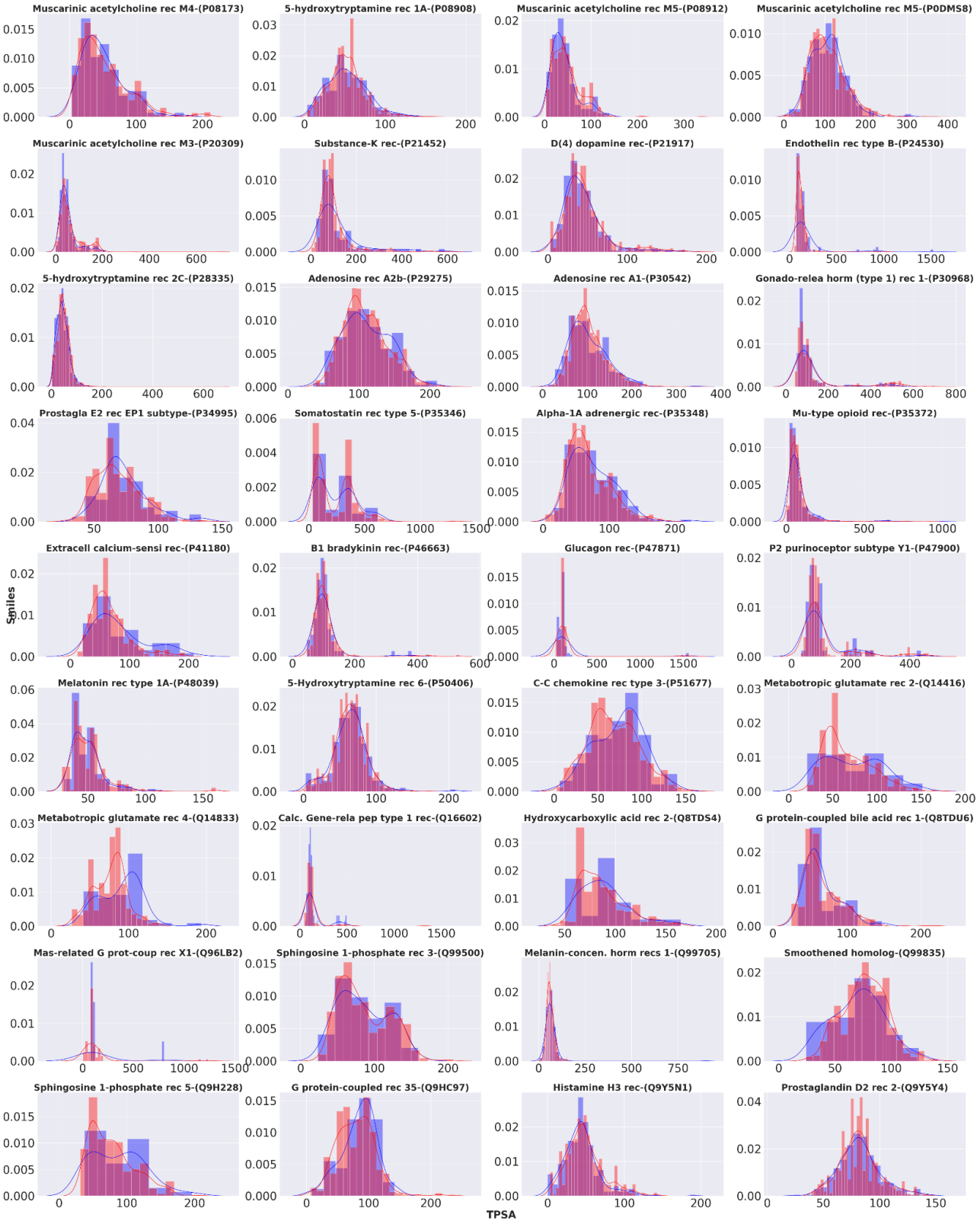


Figure A22: Histograms considering topological polar surface distribution for datasets without outliers (from cross-validation schemes) in red and considering only outliers in blue.

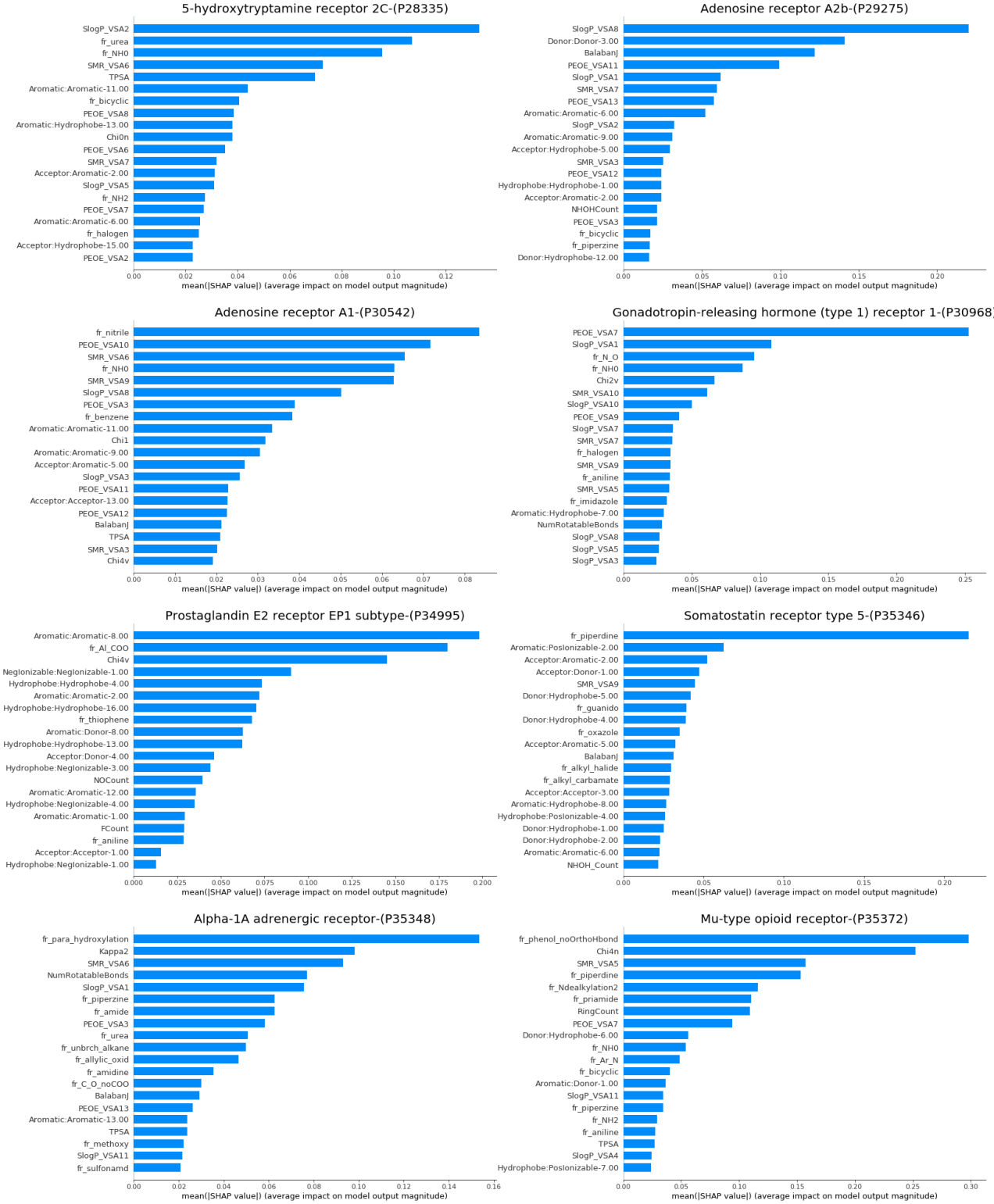


Figure A23: SHAP bar plots - Feature importance plots, the bars represent how strongly different input features affect the output of the respective model. Features are listed in descending order of importance.

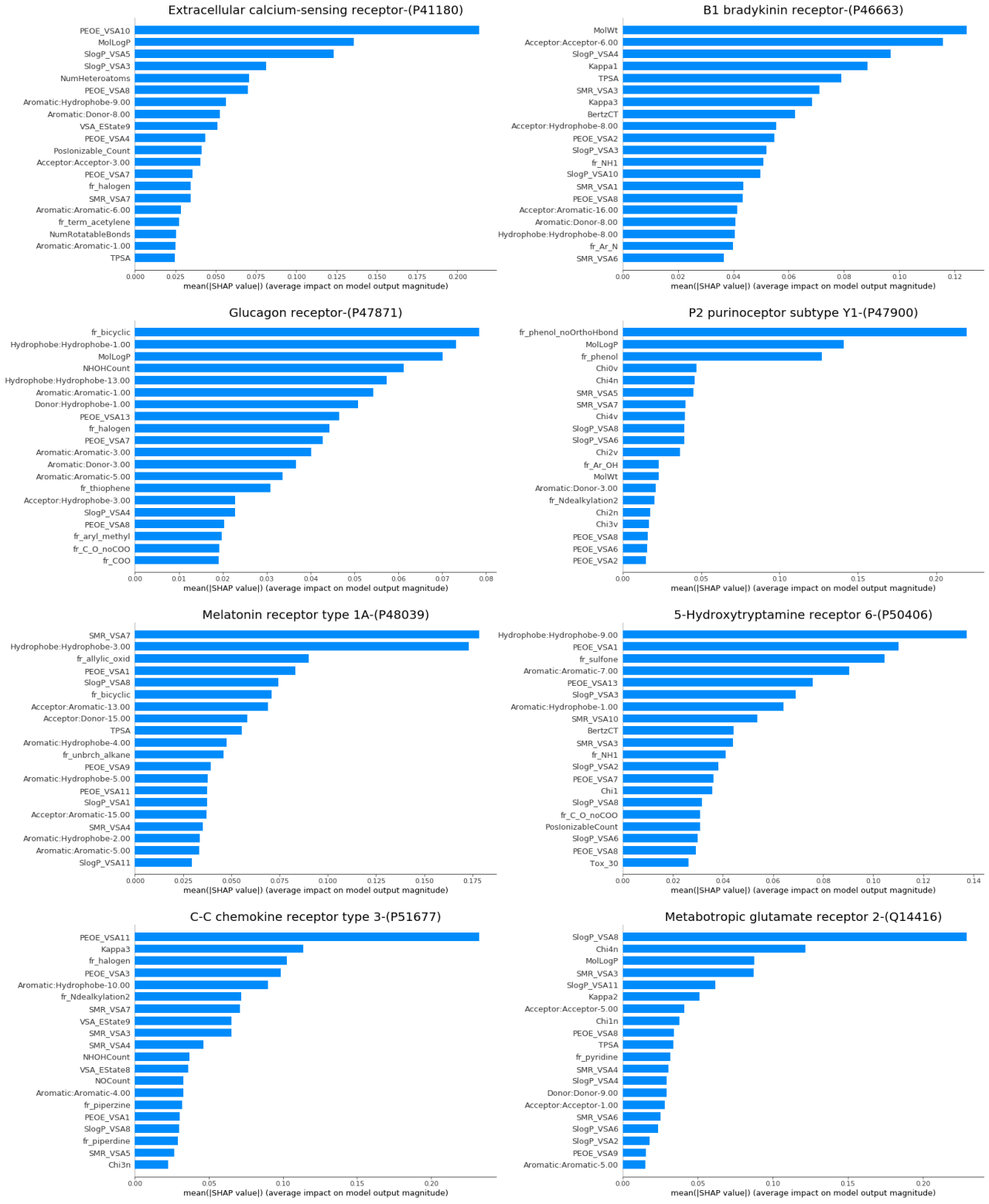


Figure A24: SHAP bar plots - Feature importance plots, the bars represent how strongly different input features affect the output of the respective model. Features are listed in descending order of importance.

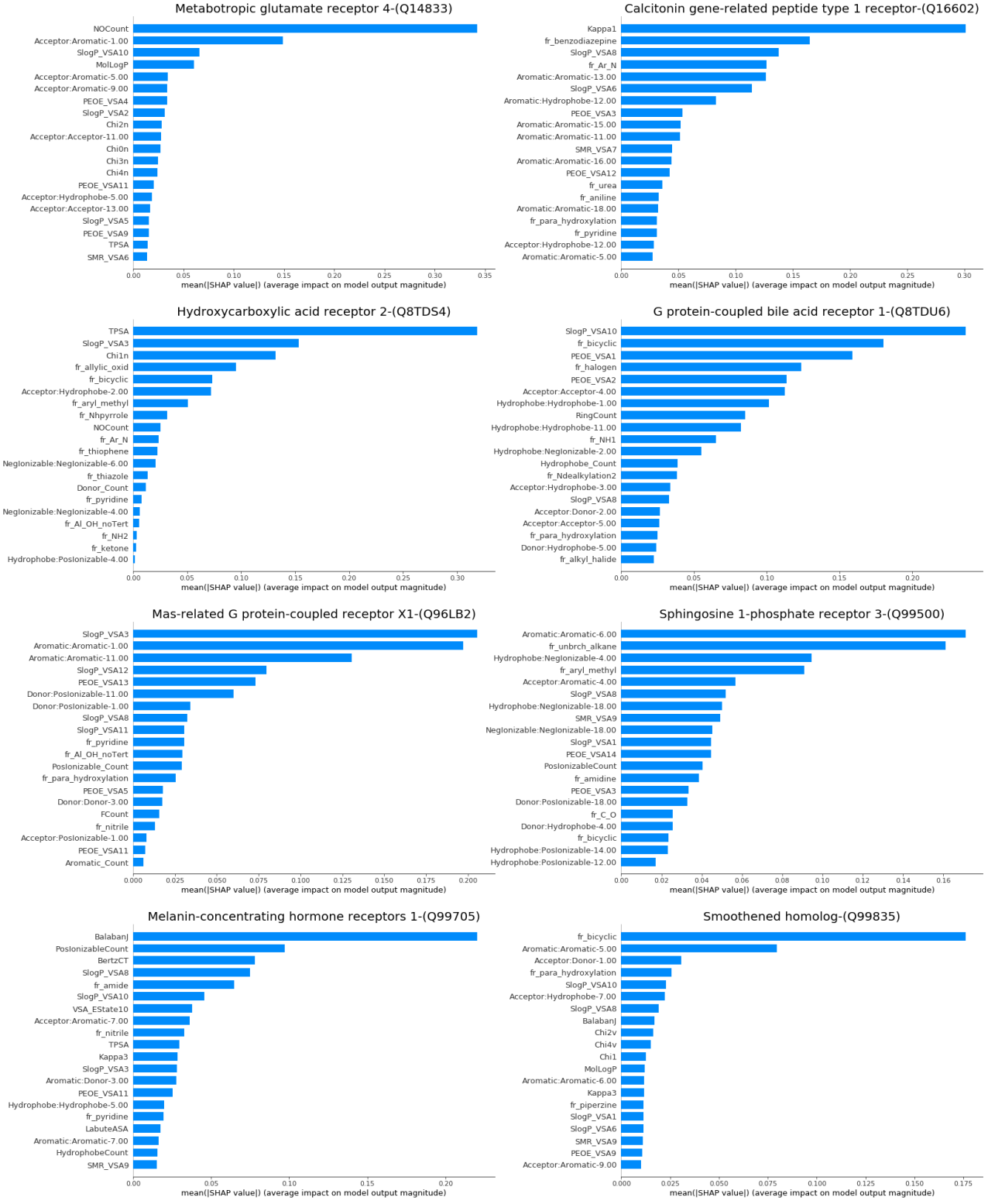


Figure A25: SHAP bar plots - Feature importance plots, the bars represent how strongly different input features affect the output of the respective model. Features are listed in descending order of importance.

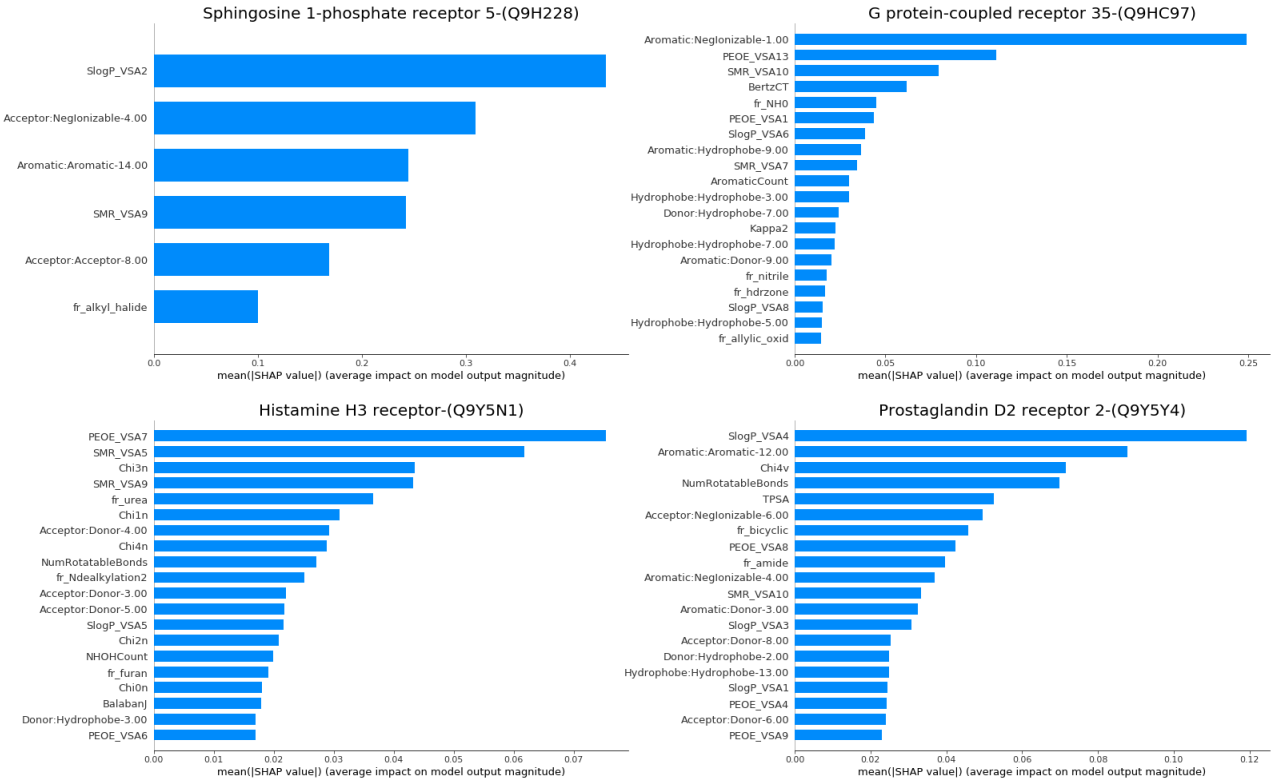


Figure A26: SHAP bar plots - Feature importance plots, the bars represent how strongly different input features affect the output of the respective model. Features are listed in descending order of importance.

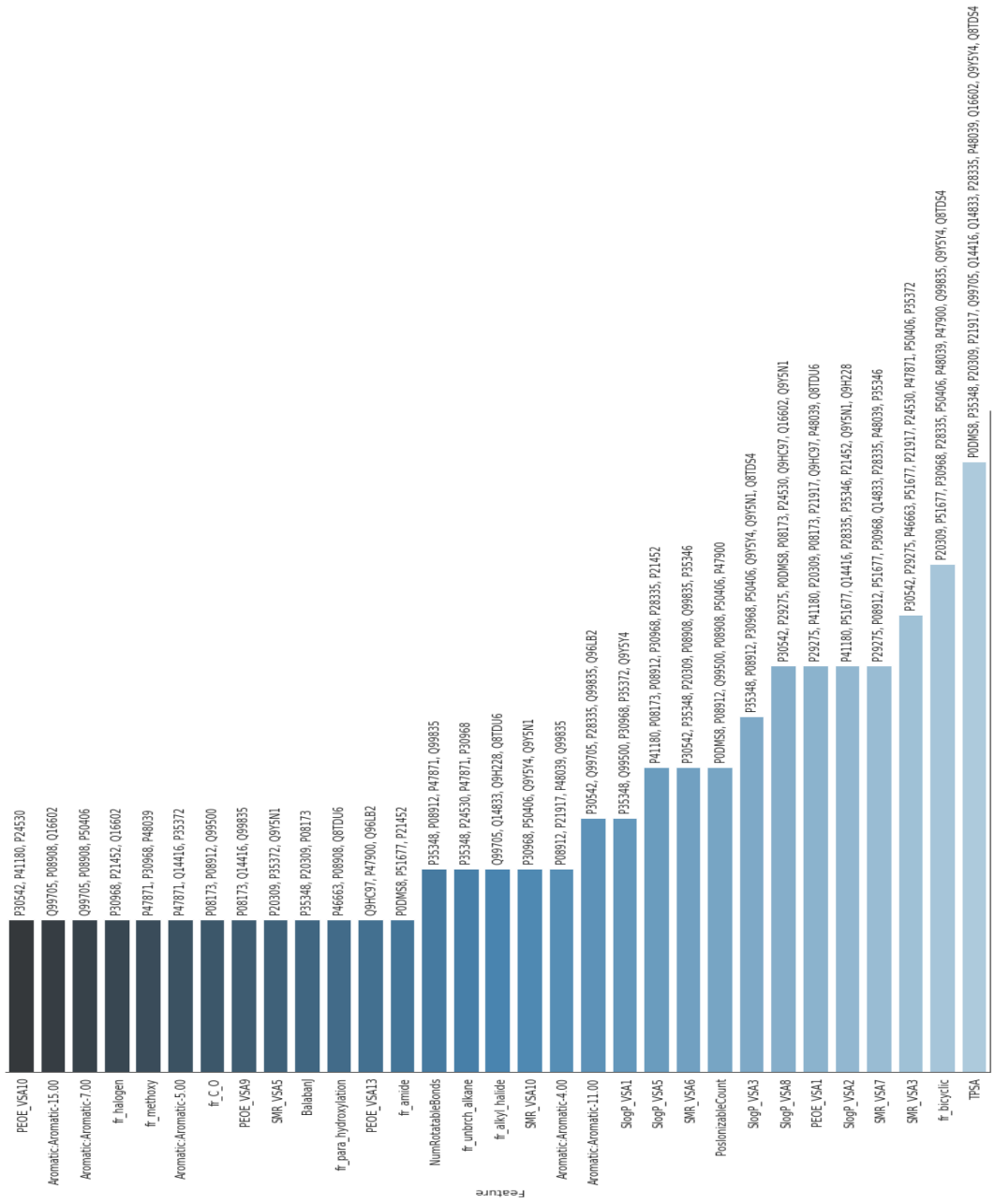


Figure A27: Distribution of the top ten features selected via forward Greedy approach for all receptors.

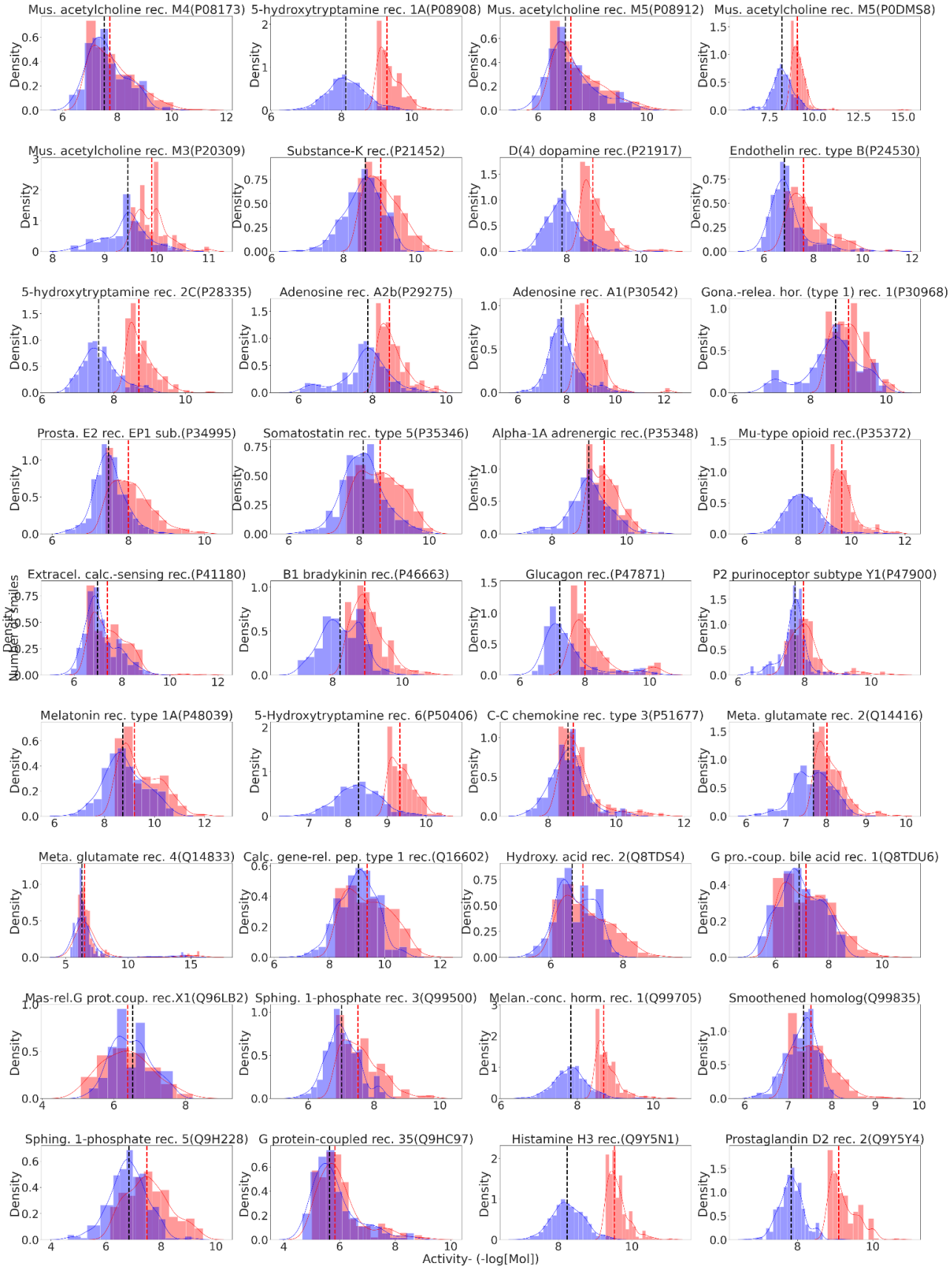


Figure A28: Histograms considering activity predicted for decoys (blue) and actual activity of the potent ligand used for generating the decoys (red). Dashed lines are the medians, decoys (black), potent ligands (red).