Evolutionary History of Calcium-Sensing Receptors Unveils Hyper/Hypocalcemia-Causing Mutations

Aylin Bircan[1], Nurdan Kuru[1], Onur Dereli[1], Berkay Selçuk[3], Ogün Adebali[1,2*],

[1] Faculty of Engineering and Natural Sciences, Sabanci University, İstanbul, Türkiye

[2] TÜBİTAK Research Institute for Fundamental Sciences, Gebze, Türkiye

[3] Department of Microbiology, The Ohio State University, Columbus, Ohio, USA.

* oadebali@sabanciuniv.edu

Abstract

Calcium-sensing receptor evolution highlights hyper/hypocalcemia-causing mutations

The Calcium Sensing Receptor (CaSR) is a key player in regulating calcium levels and has been linked to disorders like hypercalcemia and hypocalcemia. Despite advancements in understanding CaSR's structure and functions, there are still gaps in our understanding of its specific residues and their differences from receptors within the same class. In this study, we used phylogeny-based techniques to identify functionally equivalent orthologs of CaSR, predict residue significance, and compute specificity-determining position (SDP) scores to understand its evolutionary basis. The analysis revealed exceptional conservation of the CaSR subfamily, with high SDP scores being critical in receptor activation and pathogenicity. To further enhance the findings, gradient-boosting trees were applied to differentiate between gain- and loss-of-function mutations responsible for hypocalcemia and hypercalcemia. Lastly, we investigated the importance of these mutations in the context of receptor activation dynamics. In summary, through comprehensive exploration of the evolutionary history of the CaSR subfamily, coupled with innovative phylogenetic methodologies, we identified activating and inactivating residues,

24 providing valuable insights into the regulation of calcium homeostasis and its connections to
25 associated disorders.
26

27 **Introduction**

28 Calcium sensing receptor (CaSR) is a class C G-protein-coupled receptor (GPCR) that
29 maintains extracellular Ca^{2+} homeostasis by sensing calcium ions in the blood and regulating
30 parathyroid hormone release and urinary calcium (Cook *et al*, 2015; Gorvin, 2019) . The CaSR
31 is activated by Ca^{2+} and L-amino acids such as L-Phe and L-Trp as well as polyamines and
32 polypeptides (Chen *et al*, 2021; Geng *et al*, 2016; Ling *et al*, 2021). Like the other class C
33 GPCRs such as metabotropic glutamate receptors, ligands bind to the extracellular Venus
34 flytrap (VFT) domain of the receptor (Wootten *et al*, 2018).

35 Class C GPCRs are obligate dimers, forming either homo or heterodimers (Wootten *et al.*,
36 2018). CaSR forms a homodimer where each subunit is composed of an extracellular domain
37 (ECD), comprising a bilobed (LB1, LB2) VFT and a cysteine-rich domain (CRD) connected to a
38 heptahelical transmembrane (7TM) domain (Chen *et al.*, 2021; Ling *et al.*, 2021).

39 Crystal structures of the ECD (Geng *et al.*, 2016; Zhang *et al*, 2016) and cryo-electron
40 microscopy structures of the full-length CaSR (Chen *et al.*, 2021; Gao *et al*, 2021; Ling *et al.*,
41 2021; Park *et al*, 2021a; Wen *et al*, 2021) reveal the structural basis for activation mechanisms
42 and ligand binding sites. L-amino acid binding site at the interdomain cleft of LB1-LB2 (Chen *et*
43 *al.*, 2021; Geng *et al.*, 2016; Ling *et al.*, 2021; Mun *et al*, 2005; Zhang *et al*, 2014; Zhang *et al*,
44 2002) and multiple Ca^{2+} amino acid binding sites on the VFT domain are shown in the literature
45 (Chen *et al.*, 2021; Geng *et al.*, 2016; Ling *et al.*, 2021). While Ca^{2+} is the composite agonist for
46 the CaSR, L-amino acids promote receptor activation along with Ca^{2+}, but they are not able to
47 activate the receptor alone (Liu *et al*, 2020). Even though Ca^{2+} alone activates the receptor in

48 functional assays (Liu *et al.*, 2020), whether it activates the CaSR in the absence of L-amino
49 acid is still controversial (Chen *et al.*, 2021; Ling *et al.*, 2021).

50 Variants in CaSR may cause malfunctions that result in Ca^{2+} homeostasis diseases in humans.
51 More than 400 germline loss/gain-of-function (i.e., LoF and GoF, respectively) mutations cause
52 hypercalcaemic disorders, neonatal severe hyperparathyroidism (NSHPT), familial hypocalciuric
53 hypercalcemia type-1 (FHH1), and autosomal dominant hypocalcemia type-1 (ADH1),
54 respectively (Gorvin, 2019). Many more CaSR variants are anticipated to be identified as more
55 population-level genetic data become available (Gorvin, 2019). Gaining insight into the function
56 of individual residues within the receptor structure and their involvement in activation
57 mechanisms has the potential to enhance our understanding of the probability of variant
58 pathogenicity and the signaling processes of the CaSR. The examination of receptors within a
59 family and across different families allows for the identification of the specific function of each
60 residue in a receptor. However, the comprehensive understanding of the structure and
61 activation mechanisms of several families within the class C GPCRs remains elusive. This is
62 particularly true for the G-protein coupled receptor family C group 6 member A (GPRC6A) and
63 the type 1 taste receptors (TAS1Rs; specifically members 1, 2, and 3), which are the most
64 closely related subfamilies to the CaSR.

65 While all subfamily receptors of class C GPCRs share common domains and structural features,
66 the details of responding to different ligands and activating signaling pathways may differ even
67 between closely related receptors (Wootten *et al.*, 2018). Gene duplication is the main
68 mechanism that generates new protein functions across GPCRs (Flock *et al*, 2017). Protein
69 families are evolved by speciation events following gene duplication (Chagoyen *et al*, 2016;
70 Studer *et al*, 2013); thus, sequence comparisons of members within a subfamily and between
71 subfamilies can show the evolutionarily conserved domains as well as diverged protein sites
72 that distinguish one subfamily from others. One challenge with this analysis is that excessive

73 gene duplication events complicate the identification of functionally identical orthologs in a
74 subfamily. Moreover, the conservation patterns in paralogs and distant homologs may help infer
75 the specific roles of a single residue in protein function. Because the evolutionary pressure on
76 paralogs and close orthologues is not the same, substitutions allowed in paralogs may not be
77 acceptable in close orthologues. Thus, using functionally identical orthologs in sequence
78 comparisons is crucial to inferring the role of each residue in a protein family.

79 Here, we show the importance of each residue in CaSR by comparing it with the closely related
80 subfamilies, GPRC6A and TAS1Rs. We identified all orthologous sequences in each subfamily
81 by building phylogenetic trees and manually curating the duplications/speciations on the tree to
82 obtain all functionally equivalent orthologs within each subfamily. To obtain members of a
83 subfamily without requiring computationally expensive phylogenetic tree building and manual
84 curation steps, we generated highly sensitive subfamily-specific profile hidden Markov models
85 (HMMs) by using the functionally equivalent orthologs we determined using phylogenetic tree
86 analysis. We calculated a specificity score for each residue in a subfamily by calculating scores
87 based on a modified version of the PHACT algorithm (Kuru *et al*, 2022) scores which considers
88 independent evolutionary events on the phylogenetic tree while scoring the acceptability of an
89 amino acid substitution. We predicted the functional consequence of every potential substitution
90 in CaSR by using the gradient boosting trees machine learning approach. Lastly, we
91 investigated how our predictions relate to the activation mechanism of CaSR.

92

93 **Results**

94 **Evolutionary History of Class C GPCRs**

95 To reveal the evolutionary constraints on protein families, we developed a strategy to precisely
96 define a protein subfamily. A precise subfamily definition can be achieved by revealing the
97 evolutionary history of the superfamily. The evolutionary history of gene families can only be

98 established by reconstructing high-quality phylogenetic trees, which can be used to pinpoint
99 gene duplication events. Discrimination between gene duplication and speciation nodes enabled
100 us to define the paralogous and orthologous protein sequences. We further analyzed the
101 phylogenetic trees to classify the orthologous sequences that are likely equivalent in function. A
102 subfamily is defined by the human receptor and its orthologs. Within class C CPCRs, there are
103 twenty-two distinct subfamilies: CaSR, GPRC6A, three taste receptors (TAS1R1-3), eight
104 metabotropic glutamate receptors (mGluR1-8), GPRC5A-D, GABBR1, GABBR2, GPR156,
105 GPR158, and GPR179. We used functionally-equivalent orthologs in comparative analyses
106 between subfamilies, which eventually yielded subfamily-specific signatures that can be used to
107 define that particular subfamily and its function. Finally, the association between the signature
108 and function would enable a better understanding of specific molecular mechanisms and the
109 effects of variants, particularly for the protein subfamily of interest. Here, we aim to reveal the
110 signatures of the CaSR subfamily that is implied in the specific function of calcium-sensing and
111 downstream signaling.

112 We have retrieved the complete proteomes of 478 species from the NCBI database. To identify
113 proteins that belong to the class C GPCR family, we performed a hmmsearch using the seven
114 transmembrane domain profile (Pfam: 7tm_3) (Fig 1) against the proteomes. While this search
115 allowed us to retrieve the entire class C GPCRs hitting the hmm profile, it did not provide
116 subfamily annotations for the 22 subfamilies in class C GPCRs. To select canonical isoforms,
117 we performed a profile hmmscan of the PfamA profile against Class C GPCR. To generate a
118 general HMM profile for each subfamily, we first applied a BLAST search using each human
119 class C GPCR as a query (Altschul *et al*, 1990). For each subject, we blasted them against the
120 human proteome and retrieved the bidirectional best hits (Core subfamily assignment, Fig. 1).

121

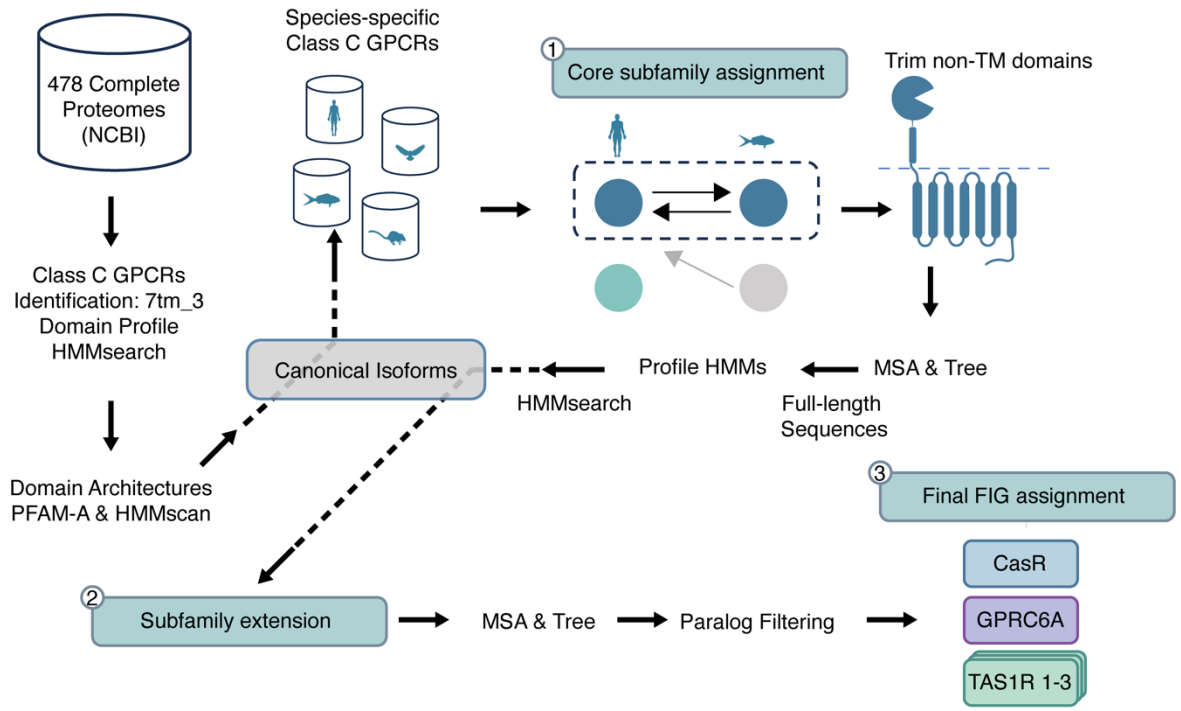


Figure 1: Summary of the Methodological Framework. 478 complete proteomes were retrieved from the NCBI database. Each sequence was searched by hmmsearch against the Pfam 7tm_3 domain profile to retrieve all class C GPCRs. Domain architectures of class C GPCRs were determined by hmmscan against Pfam. A profile to identify canonical isoforms. Species-specific BLAST databases of the canonical isoforms were built. Bi-directional mutual best hits were detected by blasting each canonical sequence against the species databases (core subfamily assignment). TM domains of core subfamily sequences were aligned, and ML trees were built to make subfamily profile HMMs. By hmmsearch against subfamily profile HMMs, other sequences in the subfamilies were found (subfamily extension). Sequences in each subfamily were aligned, and ML trees were built. Based on the ML trees, paralogs were filtered, and functionally identical groups were identified (FIG).

For proteins that did not have bidirectional mutual best hits, we assigned them to a subfamily based on their homology search against the HMM profiles generated in the previous step (Subfamily Extension, Fig 1). We produced maximum likelihood (ML) trees of extended subfamilies and filtered paralogous sequences to obtain functionally identical groups (FIGs).

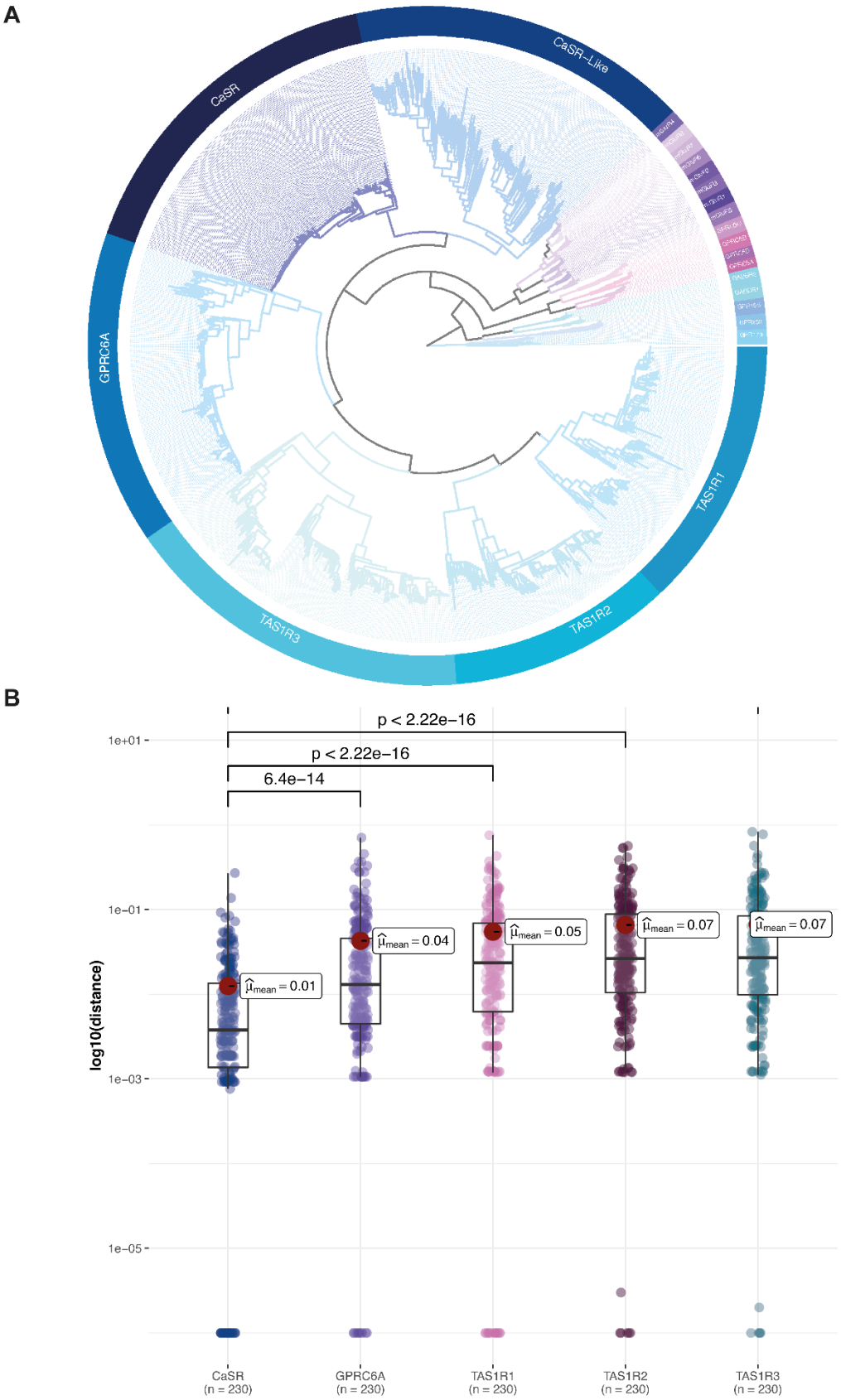
136 The CaSR subfamily produced over five thousand hits, which included vomeronasal and
137 olfactory receptors that have never been shown to sense calcium. Previous research has shown
138 that CaSR is classified in the pheromone/olfactory cluster of class C GPCRs (Pin *et al*, 2003). In
139 species that had multiple proteins assigned to the CaSR subfamily, we constructed a ML tree
140 using these hits and other human class C GPCR protein sequences. These trees revealed that
141 a significant number of duplication events occurred in the species after the clade diverged from
142 CaSR. As a result, we defined this diverged clade as a new subfamily named CaSR-likes. The
143 sequences in this subfamily are unlikely to maintain calcium homeostasis, and therefore should
144 not be annotated as calcium-sensing receptors.

145 We selected representative sequences from different species for each subfamily of 22 different
146 receptor subfamilies and 264 CaSR-like sequences and built a ML tree (Fig 2A). Also, we built
147 the ML trees of all proteins from CaSR, GPRC6A, taste receptors and merged these trees to the
148 representative tree of class C GPCRs (Fig 2A). The resulting phylogeny shows that are five
149 major clades: CaSR-related, GABA, mGluR, orphans, and retinoic acid-induced (RAIG). Orphan
150 receptors, GPR158 and GPR179, formed a clade that was diverged from other receptors
151 consistent with previous trees (Harpsoe *et al*, 2017) and had a 0.95 transfer bootstrap
152 expectation (TBE) value. γ-aminobutyric acid-B receptors (GABBR1 and GABBR2) formed
153 another clade diverging from GPR156 with 0.97 TBE. γ-aminobutyric acid-B receptorsreceptors
154 evolved earliest and have a common ancestor with the highest taxonomic rank
155 (33213−Bilateria) compared to other subfamilies. The CaSR group (CaSR, CaSR-likes,
156 GPRC6A and taste receptors) was diverged from metabotropic glutamate receptors (mGluR1-8)
157 and RAIG receptors (GPRC5A, GPRC5B, GPRC5C, GPRC5D) with 1 and 0.98 TBE values,
158 respectively. Within the CaSR group, clade CaSRs and CaSR-likes were diverged from
159 GPRC6A and taste receptors with 1 TBE. Except for TAS1R1 and TAS1R2, all CaSR group
160 subfamilies have a common ancestor from taxonomy clade 7776-Gnathostomata. TAS1R1 and

161 TAS1R2 were more specific than other CaSR group subfamilies that evolved from 117571-
162 Euteleostomi. Comparison analysis of branch lengths (Patil, 2021) among common species
163 between CaSR, GPRC6A and taste receptors shows that the CaSR subfamily is significantly
164 more conserved than its closest subfamilies (Fig 2B)

165 The higher diversity of CaSR-likes relative to CaSRs is reflected in the ML tree (Fig 2A). Branch
166 lengths of CaSR-likes are longer in contrast to shorter branch lengths in CaSR. Longer branch
167 lengths show that more variation, and thus divergence*, occurred in the CaSR-like clade.*
168 *Moreover, extensive gene duplication events occurred in this clade. For instance, rodents such*
169 *as Dipodomys ordii (taxid: 10020),* Octodon degus **(**taxid:10160**)** and snakes such as
170 Notechis scutatus (taxid: 8663) have more than a hundred receptors matching the CaSR profile.
171 However, these matches include type 2 vomeronasal receptors (V2R) and V2R-likes. Among
172 mammals, V2R genes exhibit significant variation. While dogs, cows, and primates except
173 prosimians do not have functional V2Rs**,** rodents, reptiles, and fish have multiple intact V2Rs
174 (Goes van Naters & Mucignat-Caretta, 2014). Since these receptors do not have functional
175 orthologs in mammals, they are likely to be functionally diverged, and it is crucial to separate
176 them from functionally-equivalent CaSRs***.***

177


A
CaSR
CaSR-Like
GPRC6A
TAS1R1
TAS1R2
TAS1R3
B
p < 2.22e−16
p < 2.22e−16
6.4e−14
1e+01
1e−01
1e−03
1e−05
log10(distance)
μ̂mean = 0.01
μ̂mean = 0.04
μ̂mean = 0.05
μ̂mean = 0.07
μ̂mean = 0.07
CaSR (n = 230)
GPRC6A (n = 230)
TAS1R1 (n = 230)
TAS1R2 (n = 230)
TAS1R3 (n = 230)

178

179 **Figure 2: Evolution of Class C GPCRs**. **(A)** The maximum likelihood phylogenetic tree of Class C GPCRs,
180 spanning representative species from each subfamily, is shown. Subfamilies are represented as circular layers
181 around the ML tree. All twenty-two Class C GPCR subfamilies are shown in the inner circle. In addition to these
182 subfamilies, vomeronasal and other orphan receptors are represented as CaSR-like receptors. All proteins in CaSR,
183 GPRC6A and TAS1Rs are merged into this representative species tree. **(B)** Branch lengths from leaf to root of the
184 common species that exist in all CaSR, GPRC6A and TAS1Rs are taken from the subfamily trees. Welch's t-Test
185 using the ggstatsplot package results are shown on the graph.

186

187 ***Subfamily-specific Profile HMMs to Obtain Orthologs***

188 In the class C GPCR family, gene duplication events give rise to new specificity, and each
189 duplicated gene with a new function evolves through further speciation events, producing a set
190 of orthologous sequences (Chagoyen *et al.*, 2016; Studer *et al.*, 2013). Each subfamily of class
191 C GPCRs shares a relatively conserved membrane-spanning region but also exhibits a degree
192 of variability that underlies functional differences. At the molecular level, residues that are
193 responsible for certain functional characteristics such as ligand and coupling selectivity are
194 called specificity-determining residues (Chagoyen *et al.*, 2016). Conservation analysis from
195 multiple sequence alignments (MSAs) can be used to find residues that are conserved in all
196 subfamilies through evolution as well as specificity-determining residues that are only conserved
197 in a subfamily and differ in other subfamilies. However, the success of this method depends on
198 the sequences that are used to build alignments. Therefore, it is vital to use functionally identical
199 orthologs in the analysis.

200 The seven-transmembrane domains of class C GPCRs are used to build a class-specific
201 general profile for this family (Pfam:7tm_3). However, this domain does not contain enough
202 information to differentiate subfamilies further. Moreover, excessive gene duplication events, as
203 seen in the CaSR-like clade, require precise phylogenetic analysis to differentiate between
204 CaSR and CaSR-like sequences. Also, subfamily specific profile HMMs are shown to be

promising methods to detect protein sequences belonging to a protein subfamily, as well as separation of homologs and non-homologs *(Brown et al, 2005; Srivastava et al, 2007)*. In this paper, we present a novel approach to constructing subfamily-specific profile HMMs based on the precisely produced subfamily alignments and trees. The general idea of the approach is given in Figure 3.

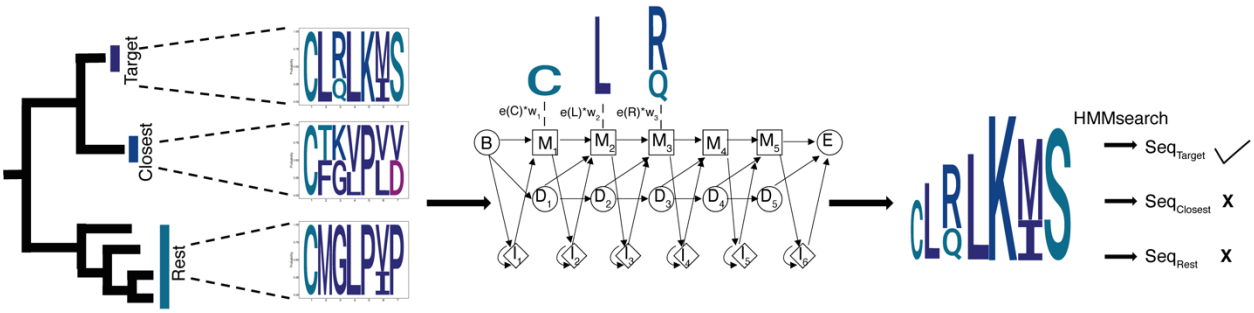


Figure 3: Subfamily Specific HMM Models. Based on the phylogenetic tree, the target, the closest, and the rest groups were determined. Initial HMMs were built without using priors. Representative amino acids in each group are selected, and their scores are calculated. According to groups, representative amino acids, and conservation scores, we calculated weights to change the emission probabilities of initial HMMs.

To construct subfamily specific profile HMMs, we first define the target family, its closest family (phylogenetic neighboring clade), and the rest based on the phylogenetic tree. We weight the identity score of each amino acid to calculate the emission probabilities. The highest weight is given to the residues which are only conserved in the target subfamily; hence, they differentiate one subfamily from the others. A minimum weight is given to the residues that are conserved both in the target subfamily and its closest clade. We tested our subfamily-specific profile HMMs' performance on independent sequences retrieved from UNIPROT (Roberts *et al*, 2019) which were not used in calculating the position weights. We assigned sequences to their corresponding subfamilies by following the same steps as the NCBI dataset (Pruitt *et al*, 2007) used to build these models. We selected new taxa that were not in the NBCI dataset to test the

226 performance of our profiles, and our subfamily specific profile HMMs correctly identified all
227 members of a subfamily while avoiding hits to proteins from other subfamilies (Table 1).

228

229 **Table 1:** Subfamily Specific Profile HMM's Performance

Subfamily HMM	Test Cases	True Positives	False Positive	False Negative	True Negative
CaSR	81	81	-	-	232
GPRC6A	62	62	-	-	251
TAS1R1	75	75	-	-	238
TAS1R2	21	21	-	-	292
TAS1R3	74	74	-	-	239

230

231 **Specificity Determining Residues**

232 CaSR is distinguished from other subfamilies of class C GPCRs by its oversensitivity to
233 substitutions that can result in either GoF or LoF mutations. This sensitivity is due to its critical
234 role in maintaining systemic calcium homeostasis and its high responsiveness to very slight
235 changes in extracellular Ca^{2+} concentrations (Gorvin, 2021). As CaSR is the most conserved
236 and ancestral subfamily among CaSR-likes, GPRC6A and TAS1Rs, it is reasonable to expect
237 that some positions can be under the relaxation of existing purifying selection in CaSR-likes,
238 GPCR6A or TAS1Rs, but not in CaSR itself. Conversely, at some positions, the same amino
239 acid may retain functional significance in both subfamilies, and at others, a position remains
240 important in each subfamily, but different amino acids are favored in each duplicate.

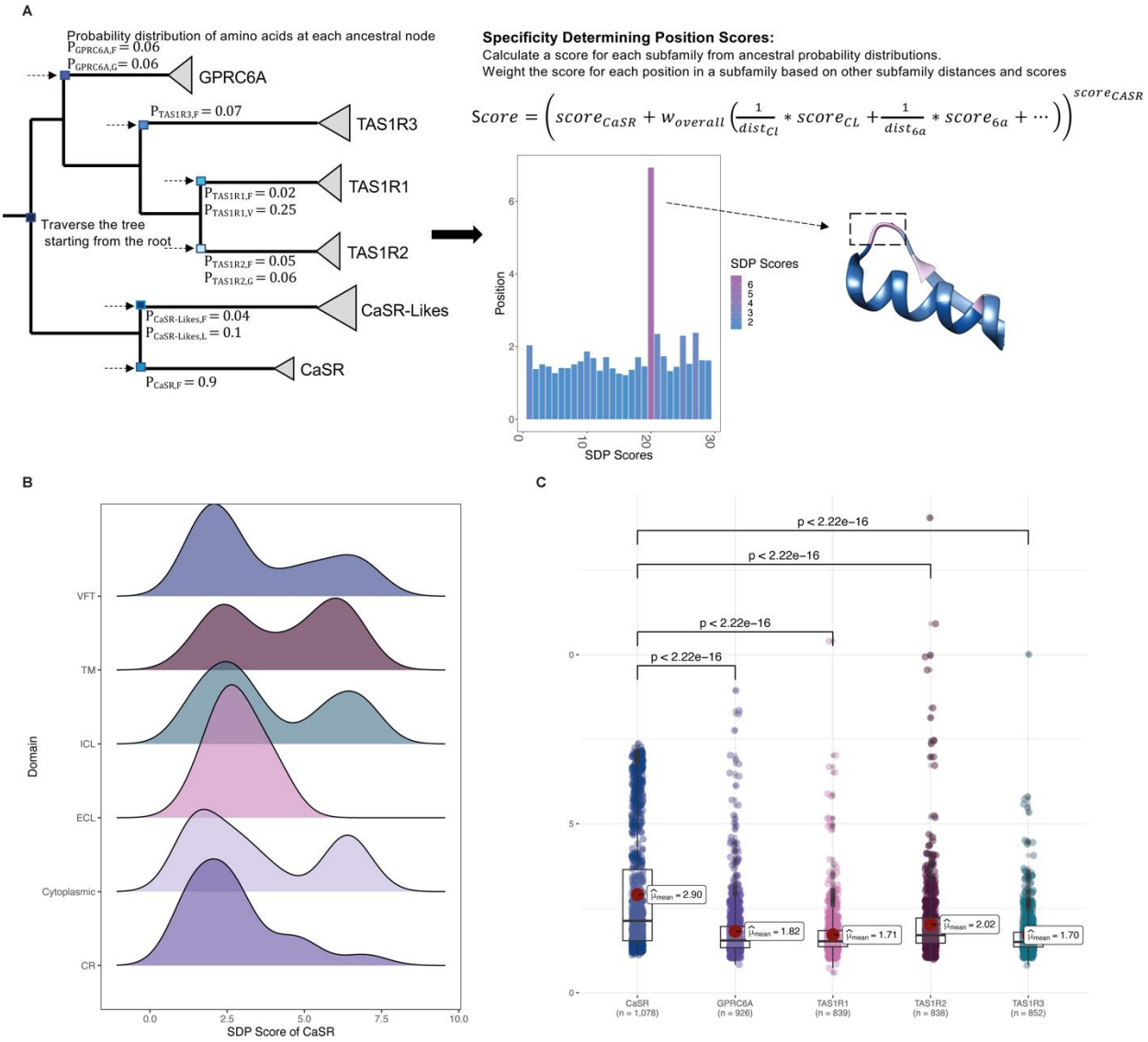
241 Specificity-determining residues that are conserved in subfamily, but differ from its sister clade
242 can be predicted by directly comparing ancestral family sequences and calculating their
243 divergence scores (Bradley & Beltrao, 2019). However, using MSAs alone does not account for
244 the number of substitution events. For example, a single substitution event in the common

245 ancestor of the bony fish clade of the CaSR subfamily can be inherited into multiple
246 descendants' sequences. Assessing this single substitution event as it repeats in each
247 sequence independently results in overcounting of these changes. Due to this mistake and
248 overcounting the effect of one single mutation repeatedly, the position is considered (i) to
249 tolerate that particular amino acid and (ii) functionally less important. In contrast, a single
250 evolutionary event might have been compensated by other substitutions in the same
251 evolutionary node. Such a substitution might not be tolerated in the other clades of the
252 subfamily.

253 Another consideration to identify and order specificity-determining residues is treating
254 substitution events on the phylogenetic tree unequally. When an amino acid in CaSR remains
255 the same but can differ in the nearby CaSR-likes subfamily, it indicates that the amino acid has
256 a CaSR-specific role. We expect the SDP score of such an amino acid to be high compared to
257 others. If an amino acid is conserved in both CaSR and remote subfamilies like taste receptors
258 but likely to be substituted in CaSR-likes, it suggests that the amino acid plays a common
259 functional role in both CaSR and other subfamilies. For such an amino acid, the SDP score
260 must be low, since it is not a specific position for CaSR.

261 *To consider these two important factors, we designed an approach to identify and prioritize*
262 residues by specificity, which differentiate a subfamily from others in the CaSR group (CaSR,
263 GPRC6A and TAS1Rs). Our approach is based on the idea presented in the functionally
264 divergent residues method (Bradley & Beltrao, 2019) along with adaption of the PHACT method
265 (Kuru *et al.*, 2022). We calculated the probability of each amino acid at each node of the CaSR-
266 group phylogenetic tree by ancestral sequence reconstruction (Fig 4A). Starting from the root of
267 the tree, we identified each substitution event and at which subfamily node that event
268 happened. After counting the number of independent substitution events in a subfamily clade
269 and comparing the probability of the same substitution in other subfamily clades, we ordered the

specificity-determining residues. We assumed that if an amino acid is allowed to change on sister subfamily nodes and is poorly conserved in sister subfamily nodes while it is highly conserved on the target subfamily node, it is a specific residue to the target subfamily only. If a substitution event is observed on a clade close to the target node, we consider that event to increase the specificity of a residue because it diverges the target group from closest sister clade. The details and the algorithm are given in materials and methods (Algorithm 3).



278 **Figure 4:Specificity Determining Position Scores. (A)** The calculation of SDP scores uses the phylogenetic tree
279 and the probability distribution of amino acids at each ancestral node. **(B)** The SDP score distributions of CaSR
280 among different domains and the SDP score distributions of each subfamily are shown. **(C)** Welch's t-Test shows that
281 CaSR has more residues with higher SDP scores compared to GPRC6A and TAS1Rs.

282 We calculated specificity scores for each CaSR, GPRC6A and TAS1Rs. Specificity score
283 distributions show that CaSR has more specific residues compared to other subfamilies (Fig
284 4B). On the VFT domain, specific residues are clustered in different regions (Fig 5). We found a
285 cluster of specific residues on the interdomain cleft between LB1-LB2 which is the L-amino acid
286 binding site in other class C GPCRs (Chen *et al.*, 2021). It suggests that this region is the
287 primary Ca^{2+} binding site in CaSR, consistent with (Liu *et al.*, 2020). We found two different
288 clusters of specific residues on the ECD. First cluster was on the LB1 domain and on the LB1-
289 LB1 dimer interface. LB1 domain plays a role in anchoring ligands and initiating domain twisting
290 by conformational changes at the interface between LB1 regions (Chen *et al.*, 2021; Ling *et al.*,
291 2021). The second cluster was found at the cytosolic side of the LB2 and at the interface
292 between LB2-CRD, where Ca^{2+} ions bind (Chen *et al.*, 2021; Geng *et al.*, 2016; Ling *et al.*,
293 2021). Interaction between LB2 subunits is required for CaSR activation that propagates to
294 large-scale transitions of the 7TMDs (Chen *et al.*, 2021; Ling *et al.*, 2021). Specific residues on
295 the LB1 domain, LB1-LB1 dimer interface, and LB2-CRD interface indicate that they provide the
296 structural conformational changes upon ligand binding to the interdomain cleft. Mutations
297 located in these regions are associated with LoF and GoF mutations (Fig 7B) (Gorvin, 2019).
298 Other specific residues are found on the CR, ECL2 and TM domains. On the ECL2 acidic
299 residues D758 and E759 are specific to CaSR. The intersubunit electrostatic repulsion between
300 the ECL2 regions could facilitate the activation of CaSR (Chen *et al.*, 2021; Ling *et al.*, 2021). In
301 the agonist+PAM bound state the ECL2 is moved by the interaction among E759, W590, and
302 K601. Deletion of D758 and E759, and single mutations of K601E and W590E disrupt CaSR
303 activity; however, Δ758–759 mutant was expressed at the cell surface with comparable levels to

304 that of WT while W590E and K601E mutants were expressed on the cell surface lower than the
305 WT level (Chen *et al.*, 2021). We found that residues W590 and K601 are not specific to CaSR.
306 The TM domains of two protomers of CaSR come into close proximity upon receptor activation
307 (Ling *et al.*, 2021).

308 The interaction between TM4-5 of each subunit in the inactive state is essential (Liu *et al.*,
309 2020), while the interaction between TM6-TM6 is crucial for the active state (Chen *et al.*, 2021;
310 Gao *et al.*, 2021; Liu *et al.*, 2020). The structural findings and the presence of CaSR specific
311 residues on each TM domain suggest that CaSR is specialized in both dimerization and ligand
312 binding. Specific residues on the TM domain are likely play a role in the regulation of
313 conformational changes observed during activation upon ligand binding and inactive states.
314 Residues inside the dimerization interface and interacting with the ligand are quite sensitive
315 substitutions because they can induce malfunctions in the receptor easily. On the other hand,
316 GPRC6A and taste receptors are more tolerant to substitutions, and they are not very
317 specialized respond to a single ion. GPRC6A and taste receptors are activated by a broad
318 spectrum of ligands (Chun *et al*, 2012; Pi *et al*, 2017). Even though the ligand of GPRC6A is
319 controversial in the literature, multiple ligands such as osteocalcin (Ocn), testosterone, basic
320 amino acids and cations such as L-Arg, L-Lys, L-Orn, calcium, magnesium, and zinc are
321 suggested to bind GPRC6A (Pi *et al.*, 2017). Taste receptors bind to different ligands, including
322 sugar, L- and D-amino acids, sweet proteins, and artificial sweeteners (Nango *et al*, 2016).

323 On the TM region, we also find the CaSR specific cholesterol recognition/interaction amino acid
324 consensus (CRAC) motif (L783, F789, S820) that is defined by the consensus (L/V)X1–5YX1–
325 5(R/K) and is often present at junctions between membrane- and cytosol-exposed domains and
326 shown in the mGluR2 receptor (Kumari *et al*, 2013). Phylogenetic analysis shows that TAS1R3
327 evolved earliest (7776 Gnathostomata) among TAS1Rs, TAS1R1and TAS1R2 subfamilies have
328 a common ancestor 117571 Euteleostomi. TAS1R3 forms heterodimers with TAS1R1 and

329 TAS1R2 (Chun *et al.*, 2012; Nango *et al.*, 2016; Nuemket *et al*, 2017). Interactions between the
330 cytosolic terminus of the extracellular CRD is needed for T1R3 dimerization. TAS1R1 and
331 TAS1R2 recognize a broad spectrum of L-amino acids that bind to the intercleft between LB1-
332 LB2 and induce the positional shift of the CRD regions; however, T1R3 loses the corresponding
333 function (Nuemket *et al.*, 2017). Our analysis showed that TAS1R1 has specific residues on the
334 LB1, LB2 and extracellular loop regions. Also, TAS1R2 has specific residues on the LB1, LB2
335 and CR domains. On the other hand, in TAS1R3, we found specific residues only on the LB1
336 and one on the CR domain. Since LB1-LB2 domains create a cavity for ligand binding, specific
337 residues on the LB1-LB2 domains of TAS1R1 and TAS1R2 may contribute to domain
338 transformation upon ligand binding. However, the number and distribution of specific amino
339 acids suggest that taste receptors are not under the same strict selective pressure as CaSR.
340

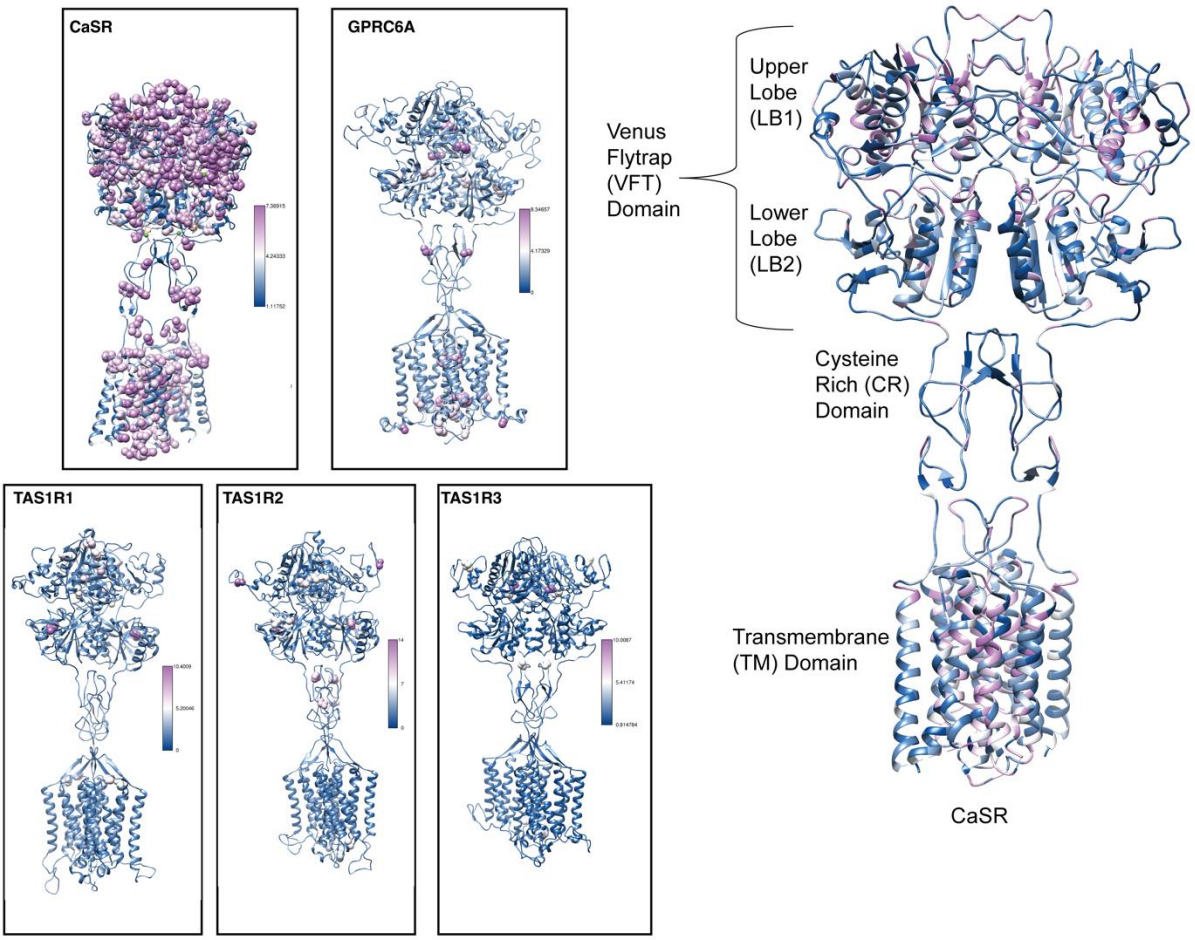


Figure 5: Specific conserved residues mapped onto structures. The cryo-EM structure of human CaSR bound with $Ca^{2+}$and L-Trp (PDB:7DTV) and homology models of GPRC6A, TAS1R1, TAS1R2 and TAS1R3 are colored based on SDP scores. Residues with a high SDP score (above 5.0) are shown as spheres. Domains on the human CaSR structure (PDB: 7DTV) are labeled and colored according to their SDP scores on the right-hand side.

Gradient Boosting Trees Machine Learning Approach to Predict the Mutation Types in CaSR

Because CaSR is a highly conserved subfamily, any substitution on the receptor disrupts the function of the receptor and causes either GoF or LoF mutations. However, predicting the functional consequences of a substitution is challenging. Evolutionary conservation of a residue

352 among subfamilies might reflect the common structural constraints, but it does not distinguish
353 between LoF and GoF mutations. In addition, at certain positions, substitution of different amino
354 acids causes either LoF or GoF mutations (Goes van Naters & Mucignat-Caretta, 2014). We
355 hypothesized that “activating” mutations are more likely to be tolerated in neighboring clades
356 such as GPRC6A and TAS1Rs and not in CaSR, whereas, in general, LoF (inactivating)
357 mutations are not tolerated in the larger clade of these receptor subfamilies. To test this
358 hypothesis and determine we can discriminate between GoF and LoF mutations in CaSR, we
359 present a tree boosting machine learning algorithm, XGBoost (Chen & Guestrin, 2016) that links
360 multiple features such as conservation scores, physico-chemical properties of amino acids, and
361 domain information.

362 Our algorithm uses sequence-based features, identity scores from MSAs, physico-chemical
363 properties of amino acids, and domain information as input features (Fig 6A). Since we
364 calculated our feature values from the MSAs, we divided our dataset into training, validation and
365 test datasets before we created feature matrices to prevent information leakage. We performed
366 50 replications with different random splittings of datasets to obtain a more robust model
367 performance.

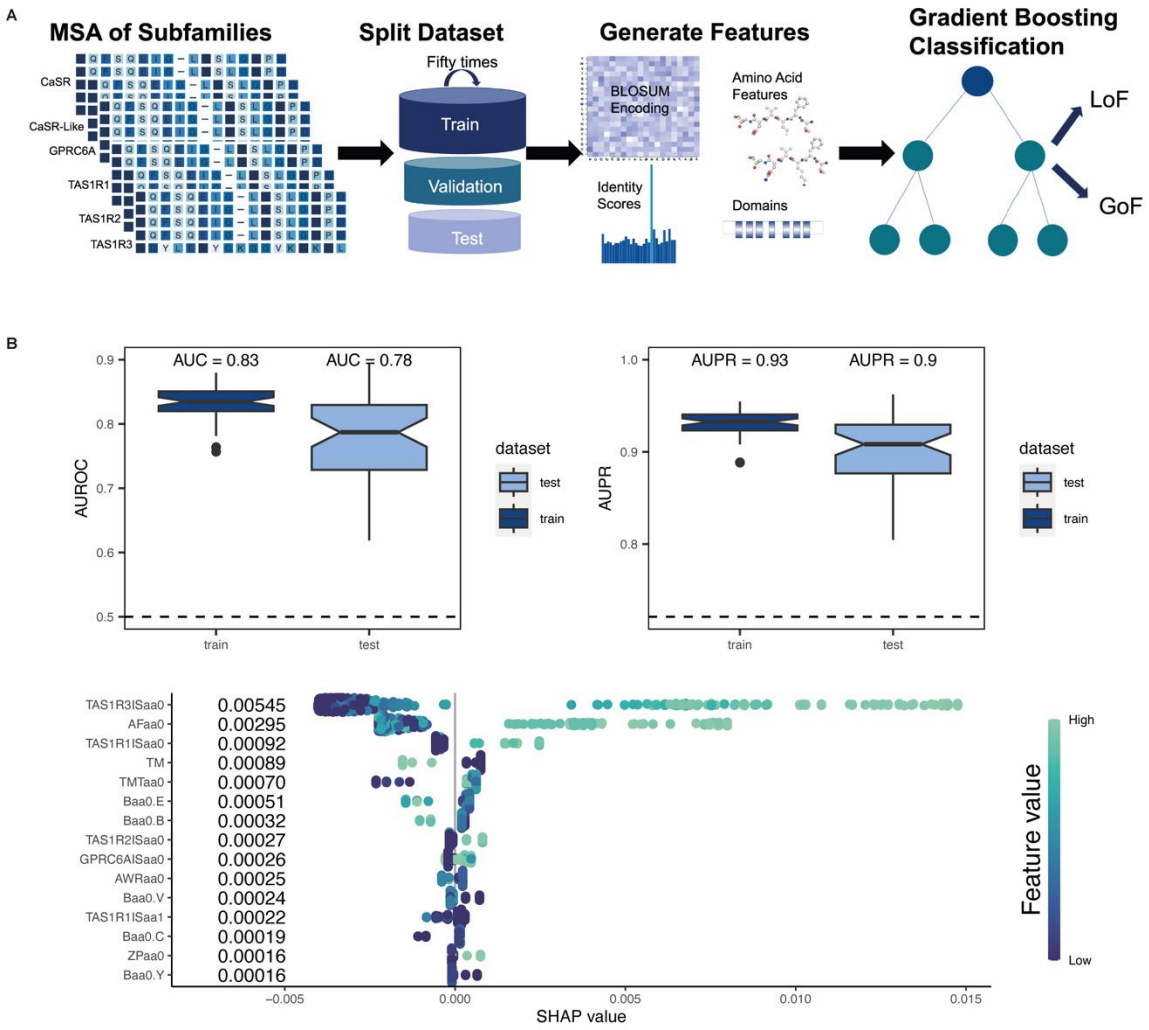


Figure 6: Gradient Boosting Trees Machine Learning Approach to Predict the Mutation Types in CaSR. **(A)**Model architecture. We used MSA of CaSR, CaSR-likes, GPRC6A and TAS1Rs to generate features as well as amino acid physico-chemical features and domain information. We performed 50 replications. **(B)** The performance and feature importance of XGBoost algorithm. The AUROC and AUPR values of 50 replications are shown. The average AUC levels of 50 replications are 0.83 and 0.78 for the train and test respectively. The average AUPR levels of 50 replications are 0.93 and 0.9 for the train and test, respectively. Contributions of Shapley values for type of pathogenicity classification to the model output for XGBoost. aa0: the amino acid found in the human CaSR, aa1: substituted amino acid, AF: average flexibility, TMT: TM tendency, ZP: Zimmerman polarity, B: BLOSUM62, AWR: atomic weight ratio, TM: transmembrane domain

Table 2: Model's predictions for the new CASR GoF and LoF mutations from literature. The correct predictions are indicated by a star symbol (*) next to them.

Mutation	Cause	Prediction
p.I857S (Chou *et al*, 2022)	hypocalcemia	gain-of-function*
p.Y825F (Moon *et al*, 2021)	hypocalcemia	gain-of-function*
p.P393R (Palmieri *et al*, 2021)	hypercalcemia	loss-of-function*
p.C60G (Li *et al*, 2021)	hypercalcemia	loss-of-function*
p.D99N (Hao *et al*, 2021)	hypercalcemia	loss-of-function*
p.T186N (Tsuji *et al*, 2021)	hypocalcemia	loss-of-function
p.A840V (Roberts *et al.*, 2019)	hypocalcemia	gain-of-function*
p.S448P (Dharmaraj *et al*, 2020)	hypercalcemia	loss-of-function*
p.L696V (Gomes *et al*, 2020)	hypocalcemia	gain-of-function*
p.D433Y (Magno *et al*, 2020)	hypercalcemia	loss-of-function*
p.S147L (Majumdar *et al*, 2020)	hypercalcemia	loss-of-function*
p.D398N (Zajickova *et al*, 2020)	hypercalcemia	loss-of-function*
p.K805R (Sagi *et al*, 2020)	hypercalcemia	gain-of-function
p.C60Y (Dong *et al*, 2020)	hypercalcemia	loss-of-function*
p.L606P (Wejaphikul *et al*, 2023)	hypercalcemia	loss-of-function*
p.H41R (Courtney *et al*, 2022)	hypercalcemia	gain-of-function
p.A110D (Bletsis *et al*, 2022)	hypercalcemia	gain-of-function
p.I139T (Wu *et al*, 2022)	hypocalcemia	gain-of-function*
p.Q164R (Coughlan *et al*, 2022)	hypercalcemia	loss-of-function*
p.T699N (Mullin *et al*, 2022)	hypercalcemia	gain-of-function
p.R701G (Mullin *et al.*, 2022)	hypercalcemia	loss-of-function*
p.T808P (Mullin *et al.*, 2022)	hypercalcemia	loss-of-function*

The ROC and PR curves are used to understand the performance of a binary classifier that assigns each element of data into two groups. The ROC curve is a graphical plot that shows the false positive rate versus the true positive rate for different threshold values between 0 and 1. A PR curve is a plot of the precision and the recall for different threshold values, and it is useful for imbalanced datasets. We used the areas under the ROC and PR curves (i.e., AUC and AUPR,

386 respectively) to compare the performances of the model on the train and test datasets for 50
387 replications. Higher AUC and AUPR values are associated with better performance. AUC and
388 AUPR over all replications were shown in (Fig 6B). Our average AUC values for training and
389 test among 50 replications are 0.83 and 0.78 (Fig 6B). Our average main AUPR values for
390 training and test among 50 replications are 0.93 and 0.9, respectively (Fig 6B). After we
391 reported our algorithm performance, we trained our algorithm with the whole dataset. We tested
392 our algorithm with new test cases from the literature (Table 2). Additionally, we categorized
393 amino acids that are observed in the CaSR MSA as neutrals. To date, no pathogenic
394 substitution has been reported in the literature for these amino acids that we identified as
395 neutral. We visualized all predictions in the form of a heatmap for every other amino acid at
396 each position until the disordered region (position 892) of the human CaSR (Fig 7A). We
397 mapped known CaSR LoF and GoF mutations on the cryo-EM structure of human CaSR bound
398 with Ca^{2+} and L-Trp (PDB:7DTV)(Ling *et al.*, 2021) (Fig 7B). There is a tendency that GoF
399 mutations are in the inner-core regions. In the heatmap, we observed a similar prediction
400 pattern: GoF predictions are mostly in the inner-core regions (Fig 7A). SHAP (SHapley Additive
401 exPlanations) values provide a way to decode the inner workings of a machine learning model
402 like XGBoost. These values calculate the average contribution of each feature to the overall
403 prediction, taking into account any interactions between the features. Based on the SHAP
404 values, the conservation scores of human CaSR amino acids in other subfamilies play a
405 significant role in the model's prediction, as shown in Figure 7B. If the amino acid is also
406 conserved in GPRC6A and taste receptors (in fact, conservation score in TAS1R3 has the
407 highest contribution), the model predicts a substitution of that amino acid as LoF. On the
408 contrary, when an amino acid is conserved exclusively in CaSR, substituting that amino acid is
409 predicted to result in a GoF. New test cases (Y825F, A840V, L696V, I139T) mentioned in Table 2
410 have been accurately predicted to cause GoF mutations. These specific amino acids are
411 conserved either solely in CaSR or in GPRC6A and CaSR-like receptors, but not across all

412 receptor types. Notably, the A840V substitution leads to a GoF mutation, as valine is conserved
413 in GPRC6A. These analyses reinforce our hypothesis that residues conserved in CaSR, but not
414 in other subfamilies, are more likely to induce GoF mutations. Additionally, amino acids
415 conserved within closely related subfamilies might disrupt the functioning of other subfamilies.
416 On the other hand, the new LoF cases shown in Table 2 (D99N, C60G, Q164R, T808P) show
417 that reference amino acids are conserved across subfamilies.

418 Another important feature is the structural location of the amino acid. Our findings indicate that if
419 the amino acid is located in the TM domain, a substitution would result in a GoF mutation. It is
420 known that the majority of GoF mutations are located in the TM domain, as shown in Figure 7B.
421 The presence of certain amino acids on the TM domain of CaSR suggests that they play a
422 crucial role in its activation mechanism. Even though substituting those amino acids might be
423 acceptable in GPRC6A and taste receptors, they might lead to the lock of TM domains and
424 result in the overactivation of CaSR.

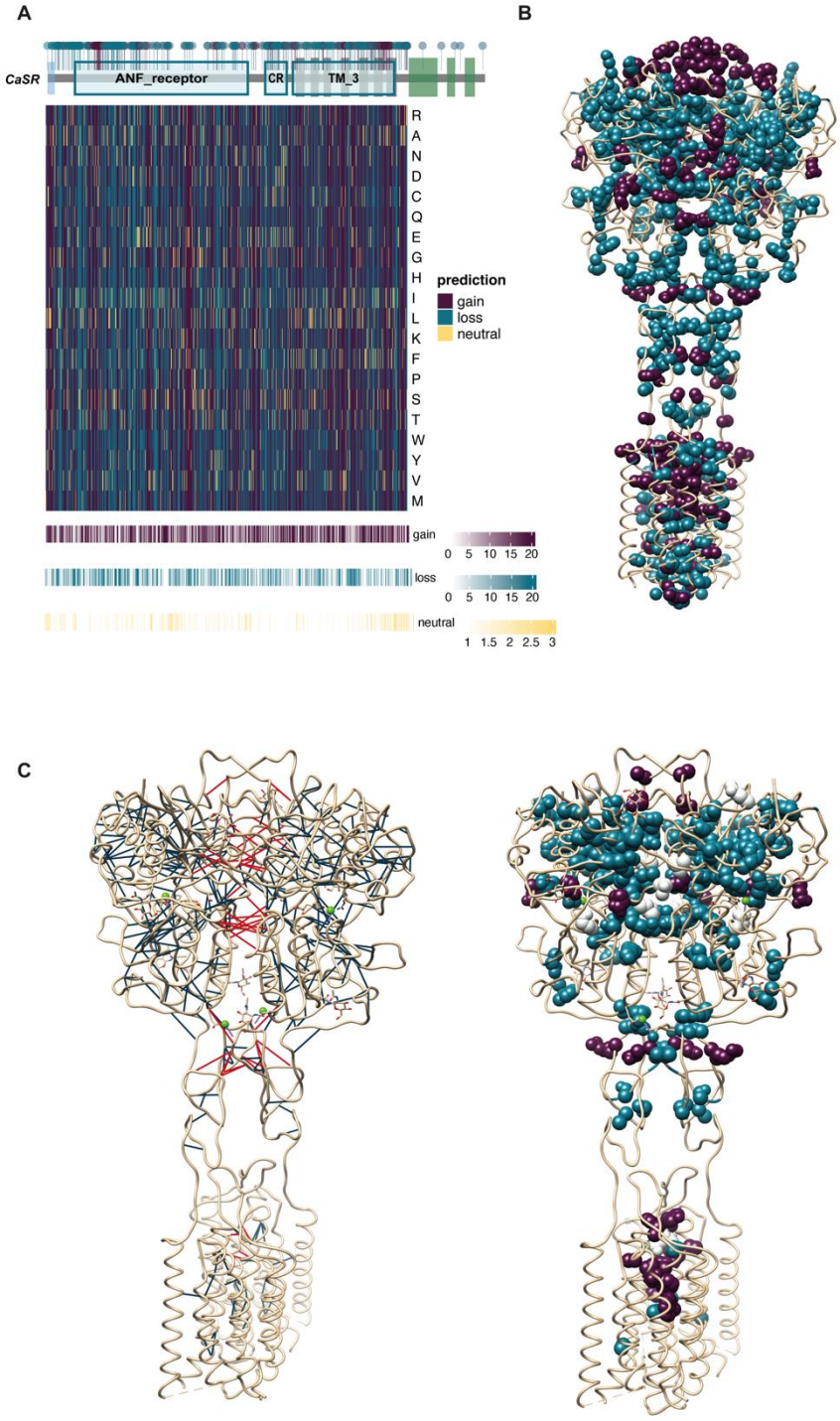
425 To gain insight into the activation mechanism of the CaSR and its association with activating
426 and inactivating mutations, we analyzed the complex network of residue-residue connections by
427 comparing active state CaSR structures with inactive state structures. For the monomer
428 activation network, we modified the residue-residue contact score (RRCS) algorithm (Zhou *et al*,
429 2019) to detect atom level contact changes in identical amino acids in different states. We made
430 this change for two main reasons. First, class C GPCRs do not exhibit large structural changes
431 during receptor activation (Hauser *et al*, 2021) , like other GPCR classes. Therefore, residue
432 level changes do not provide enough resolution for understanding the mechanism, especially for
433 the TM domain. Second, residue level changes can only highlight certain important positions
434 while investigating atom level contact changes enables us to understand the impact of more
435 residues. For the changes observed within the dimerization interface, we used the original
436 algorithm. To build the networks, we identified significant contact changes observed in two

437 receptor states during activation both in the individual protomer and at dimer interface. (Fig 7C).
438 Significantly, the residues that were linked to changes resulting in increased activation were
439 more abundant in critical locations such as the loop of lobe 1 inside the VFT domain, the
440 interdomain cleft, the CR domain, and the TM domains (Fig 7C).

441 Recent studies have highlighted the role of domain twisting in the CaSR homodimer's activation
442 process, initiated by intersubunit domain contacts and conformational shifts at the interface
443 between the lobe 1 regions (Ling et al., 2021). In this study, we have noticed a notable
444 augmentation in the level of contact between two pivotal amino acid residues, L125 and Y20,
445 located within the loop region of LB1 domains in the homodimeric structure. Notably, the
446 substitutions of L125F and L125P led to mutants that exhibited a GoF phenotype. In the TM6
447 domain, we observed increased interaction between residues in the dimer, with particular
448 attention to A824 and P823. Significantly, residue A824, which exhibits specific conservation,
449 has been associated with GoF mutations in the A824P and A824S (Gorvin, 2019)

450 We identified a notable increase in the level of contacts observed between residues within the
451 CR domain. Specifically, these interactions have been observed between two subunits of the
452 dimer and involve the following interactions: E556-D578, E556-S580, and E556-K552. Although
453 only residue S580 demonstrated specificity for CaSR, other acidic residues either exhibited
454 tolerance towards replacements with other acidic residues or remained conserved across all
455 subfamilies of CaSR. Nevertheless, it is important to acknowledge that the substitution of E556K
456 resulted in a mutation with enhanced activation. The results indicate that the interactions
457 between these particular residues are of utmost importance in maintaining the receptor's active
458 conformational state. This conclusion is consistent with previous structural investigations that
459 have demonstrated the convergence of LB2, CR, and TM domains in both subunits during the
460 twisting of CaSR, leading to a more condensed conformation in the active state of CaSR (Chen
461 et al., 2021; Ling et al., 2021). In contrast, we have also observed distinct residues linked to LoF

462 mutations. These residues exhibit enhanced interactions within a single subunit, predominantly
463 located in the LB1 and LB2 sections of the VFT domain, specifically in proximity to the Ca^{2+}
464 binding sites (Fig 7C). Residues that induce LoF upon mutation are primarily located within the
465 core of the VFT domain. This implies that any modification in amino acids could potentially
466 induce structural alterations, ultimately resulting in misfolding or disruption of the activation
467 mechanism and consequent LoF.



468

469 **Figure 7: LoF and GoF Mutation Predictions. (A)**Visualizing the results of our XGBoost model. The heatmap
470 displays the XGBoost model's predictions for each of the 20 amino acids at every position except disordered regions
471 (892-1078) within the human CaSR. Above the heatmap, the domains of the CaSR are shown. Within these domains,

472 circles represent all known LoF and GoF mutations documented in the literature. Circles denoting GoF mutations are
473 colored purple, while those representing LoF mutations are colored blue. Below the heatmap bar graphs show the
474 number of GoF, LoF and neutral predictions among the 19 possible substitutions. **(B)** Mutations on the human CaSR
475 structure. LoF- and GoF-associated mutations are shown on the cryo-EM structure of human CaSR bound with Ca^{2+}
476 and L-Trp (PDB:7DTV) as blue and red spheres, respectively. **(C)** Increased residue-residue contacts are shown on
477 the cryo-EM structure of human CaSR bound with Ca^{2+} and L-Trp (PDB:7DTV) on the left. Interdomain and
478 intrasubunit interactions are shown as red and blue lines, respectively on the right. LoF- and GoF-associated
479 mutations among the interacted residues are shown as blue and purple spheres, respectively. Switch residues are
480 shown as white spheres.

481

482 **Discussion**

483 In this study, we showed the evolution of CaSR by developing a methodology for precisely
484 defining functionally equivalent orthologous sequences across species and therefore
485 subfamilies. We built a high-quality phylogenetic tree of CaSR with its closest subfamilies,
486 GPRC6A and TAS1Rs. Statistical analysis of branch length distances from this phylogenetic
487 tree showed that CaSR is evolutionarily more conserved compared to GPRC6A and TAS1Rs.
488 While GPRC6A and taste receptors can bind to a diverse range of ligands and are able to
489 tolerate substitutions at most of the positions, CaSR requires a delicate balance for proper
490 functioning.

491 The high evolutionary conservation and specificity of CaSR in contrast to the closest subfamilies
492 are reflected in the SDP score analysis. CaSR has specific residues clustered in different
493 regions of the receptor. They are located on Ca^{2+} and L-Trp binding sites on the VFT, as well as
494 on the dimerization sites between two sub-units of the homodimer. Specific residues on the
495 dimer interfaces indicate that dimerization maintained by interactions between different subunits
496 is required for ligand binding and the correct activation of the CaSR. Ca^{2+} ion binding and
497 interactions between LB2-CR domains and conformational changes in LB1 domain were

498 suggested to be required to activate CaSR (Chen *et al.*, 2021; Geng *et al.*, 2016; Ling *et al.*,
499 2021). Mutational analysis at some positions on the LB1 domain has been shown to reduce the
500 effect of Ca^{2+}-stimulated intracellular Ca^{2+} mobilization in cells (Chen *et al.*, 2021; Ling *et al.*,
501 2021). In contrast, substitutions caused negative charge neutralization on the ECL2 result in
502 prompting the activation of CaSR (Ling *et al.*, 2021). Our results suggest that residues with low
503 SDP scores on any domain are required for a common activation mechanism since they are
504 conserved across functionally different receptor subfamilies. However, residues with high SDP
505 scores cause malfunctions in the CaSR. Any substitution in a residue with a high SDP score
506 might either cause over or less activation. Deep mutational scanning approaches or new
507 methods that simultaneously profile variant libraries (Jones *et al*, 2020) are needed to provide
508 further evidence to functionally assay all possible missense mutants.

509 To predict the functional consequence of a mutation in human CaSR, we used the Extreme
510 Gradient Boosting (XGBoost) method. XGBoost is able to perform well on small datasets by
511 incorporating a variety of regularization methods to control the model complexity, which helps to
512 prevent overfitting. We have a small and unbalanced dataset in that the number of GoF
513 mutations was very low, therefore it is prone to overfit. To prevent overfitting while achieving
514 high predictive performance, we used a simple method along with regularization parameters.
515 Moreover, we tried to keep the ratio between the number of LoF and the number of GoF
516 mutations for training and test sets as close as possible. To get a robust performance, we
517 repeated the train-validation-test splitting procedure fifty times. To increase predictive
518 performance, we could use more complex methods, such as deep learning, but they require
519 larger datasets. Studies that used deep learning or ensemble methods for similar assessments
520 are different in terms of prediction, in which they predict the type of mutation as either
521 pathogenic or neutral (Alirezaie *et al*, 2018; Ioannidis *et al*, 2016; Rentzsch *et al*, 2019; Rogers

522 *et al*, 2018). Even though there are a number of mutations of human CaSR in the Clinvar, the
523 functional consequences of most of them are not known.

524 Given the constraints of the small dataset and limited additional data, we carefully selected and
525 processed the features for our model's training. Features that are used to train a machine
526 learning model heavily determine its performance. The more features we use, the more
527 information the model has to learn from, which can lead to improved predictive performance.
528 However, having too many features can also lead to overfitting. Moreover, the quality of the
529 features is more important than the quantity. One important evolutionary process that can affect
530 the functional consequences of a substitution is co-evolution. From the MSA of CaSR proteins,
531 we manually selected six positions, p.180,p.212,p.228,p.241,p.557 and p.883, that are in our
532 dataset and co-evolved. We masked the co-evolved amino acids from the MSA and performed
533 the train-validation-test splitting procedure fifty times again. Our average AUC values for training
534 and test among 50 replications were 0.83 and 0.77, respectively and our average AUPR values
535 were 0.93 and 0.89. Despite not experiencing an improvement in performance, we found that
536 the amino acid changes p.I212T, p.F180C, and p.I212S were now predicted to cause LoF,
537 contrary to their previous prediction of causing GoF. We cannot accurately assess the impact of
538 co-evolution on performance because there is a lack of effective tools for identifying co-evolved
539 positions and our dataset contains only a limited number of co-evolved positions, but we
540 anticipate that it is an important feature to differentiate GoF and LoF mutations.

541 We built subfamily-specific profile HMMs to get all functionally-equivalent orthologs while
542 excluding other proteins. To generate these HMM models, we manually decided target, closest
543 and rest groups based on the phylogenetic tree of CaSR group. Based on the nature of a
544 phylogenetic tree, the selection of these groups is changed, so that this process can be further
545 automated. We did not anticipate that our specific models would match any receptors from other
546 classes of GPCRs, since they are evolutionarily more distant to CaSR group. We expect that

547 our subfamily specific profile HMMs can be used to obtain orthologs in different protein families
548 for the upcoming genomes. They can be particularly useful for studying protein families with
549 many duplications and orphan protein families, where it can be difficult to identify true members.
550 These models are particularly important to avoid computationally expensive and expertise-
551 required phylogenetic tree reconstruction and analysis.

552

553 **Materials and methods**

554 **Class C Proteins and Their Domain Architectures**

555 478 complete eukaryotic proteomes were downloaded from the NCBI genomes website
556 (https://ftp.ncbi.nlm.nih.gov/genomes/archive/old ref seq/) in 2018. A hmmsearch of HMMER
557 software (Eddy, 2011) (http://hmmer.org/) was run for each proteome against the Pfam 7tm_3
558 profile (Finn *et al*, 2016). Sequences with significant 7tm_3 hit based on hmmsearch results
559 (above the default threshold) were compiled from proteomes. A hmmscan of HMMER software
560 (Eddy, 2011) (http://hmmer.org/) was run for these sequences against the Pfam-A 32.0
561 database (Finn *et al.*, 2016). Based on the results of the hmmscan, the longest isoform was
562 taken and saved in a separate file named by taxonomic ID, however, canonical sequences were
563 obtained for human (based on the given canonical proteins on the UniProt website (Consortium,
564 2017). Because plants do not have GPCRs, they were eliminated from the analysis. For single
565 isoform sequences of each proteome, a BLAST database was built (Altschul *et al.*, 1990).

566

567 **Subfamily Definition and Subfamily Specific Models**

568 Each protein sequence of each taxon was queried through BLASTP against each prepared
569 BLAST database (Altschul *et al.*, 1990). reciprocal mutual best hits of each human class C
570 GPCR were collected in a file named gene id. reciprocal mutual best hits of each class C GPCR
571 and remaining human class C GPCRs were collected and 7TM domains of these sequences

572 were taken based on hmmscan results (the longest sequence that hit the 7tm_3). Sequences
573 were aligned using the MAFFT v7.221 E-INS-I algorithm with default parameters (Yamada *et al*,
574 2016). A maximum likelihood based phylogenetic tree of each subfamily of class C GPCR was
575 built using RAxML version 8.2.12 with automatic protein substitution model selection
576 (PROTGAMMAAUTO) and 100 rapid bootstrapping parameters (Stamatakis, 2014). The most
577 common lowest taxonomic level was added to the phylogenetic tree with the ETE toolkit
578 (Huerta-Cepas *et al*, 2016). Based on the phylogenetic tree, sequences belonging to the
579 corresponding subfamily were taken, and a profile HMM was built. Subfamily Assignment The
580 process begins by scanning each sequence with a 7tm_3 domain against profile Hidden Markov
581 Models (profile HMMs). After the sequence is scanned, the subfamily is determined based on
582 three conditions: (1) The maximum score value of the hmmscan must belong to the given
583 subfamily. (2) E-value is a measure of the significance of a match in a database search, and the
584 lower the E-value, the more significant the match is. The E-value of the sequence must be the
585 lowest. (3) The sequence must belong to the most common highest taxonomic level of the given
586 subfamily. Taxonomic level refers to the classification of an organism within a biological
587 classification system. If a sequence meets these three conditions, it is assigned to the
588 corresponding subfamily. After this, the full length sequences of each subfamily were then
589 aligned using the MAFFT v7.221 algorithm (Yamada *et al.*, 2016) and trimmed using the gappy-
590 out method of the trimAl tool (Capella-Gutierrez *et al*, 2009).

591

592 **Paralog Filter**

593 There were a number of duplications in the CaSR subfamily. For example, Dipodomys ordii has
594 116 CaSR sequences. To reduce the number of sequences, human CaSR and other human
595 class C GPCR proteins sequences compiled with CaSR sequences of each taxon and aligned
596 with the MAFFT v7.221 auto algorithm (Yamada *et al.*, 2016), and the gappy-out method of the

597 trimAl tool was used to trim the MSA (Capella-Gutierrez *et al.*, 2009). The ML tree was built
598 using RAxML-NG v0.9.0 with ML tree search and bootstrapping (Felsenstein Bootstrap and
599 Transfer Bootstrap) parameters (Kozlov *et al*, 2019). Based on the ML tree, proteins that were
600 diverged from the common ancestor of the human CaSR clade were classified as CaSR-likes.
601 Proteins that were clustered with the human CaSR were accepted as CaSRs. After we assigned
602 all proteins to their subfamilies, we built final ML trees for CaSR, GPRC6A, and TAS1Rs. We
603 added human CaSR sequence to the GPRC6A and TAS1Rs subfamilies, and human GPRC6A
604 sequence was added to CaSR subfamily as an outgroup. We aligned subfamily sequences with
605 the MAFFT v7.221 einsi algorithm (Yamada *et al.*, 2016) and built the ML trees by using
606 RAxML-NG v0.9.0 with the FTT model parameter (Kozlov *et al.*, 2019). We labeled the
607 duplications at each node on the ML trees. Based on the duplications, we manually checked the
608 trees and removed a clade that was a subset of its sister clade by using the ETE toolkit (Huerta-
609 Cepas *et al.*, 2016). We took each branch and node length from leaf to root of the tree by using
610 common species in all CaSR, GPRC6A and taste receptor trees to calculate subfamily
611 conservation by using Welch's t-Test by using the 'ggstatsplot' package (Patil, 2021).
612

613 **Subfamily Specific Profile HMMs**

614 After we took all receptors from CaSR, CaSR-like, GPRC6A, and taste receptors, we aligned
615 them by using MAFFT v7.221 auto algorithm (Yamada *et al.*, 2016). For each subfamily we
616 removed the positions from the MSA that correspond to a gap in the human receptor. Then, we
617 divided the MSA into subfamily alignments. We generated an HMM profile from the gap-
618 removed alignment of each subfamily. The "pnone" option of the HMMER package was used in
619 the HMM profile construction step, and thus the obtained probability parameters are simply
620 calculated by employing observed frequencies. The constructed HMM profiles for any target
621 subfamily, even with the "pnone" option, still hit the sequences from other subfamilies. On the

622 other hand, we aim to obtain profiles such that they can correctly identify the elements of the
623 target subfamily while not hitting any proteins from the remaining ones. To achieve this, we use
624 a position-weighting approach. After determining the weight per position, we update the
625 emission probabilities by taking these weights into account. The weights are computed by a
626 detailed analysis of the conservation dynamics of each subfamily. As the initial step, we defined
627 the target, the closest and the rest groups based on the ML tree. By considering the distance
628 between the root node of each subfamily and the target one, we determined the closest group.
629 The remaining parts of the tree were taken as the rest. To obtain specific HMM profiles for each
630 subfamily, we considered five different scenarios that show variety in terms of the subfamilies in
631 the target, close and rest groups.

632 - CaSR is the target group, CaSR-likes are the closest group, and GPRC6A and taste
633 receptors (TAS1Rs) are the rest.
634 - GPRC6A is the target group, TAS1Rs are the closest group, CaSR and CaSR-likes are
635 the rest.
636 - TAS1R1 is the target group, TAS1R2 is the closest group, and TAS1R3 is the rest.
637 - TAS1R2 is the target group, TAS1R1 is the closest group, and TAS1R3 is the rest.
638 - TAS1R3 is the target group, TAS1R1 and TAS1R2 are the closest group and GPRC6A
639 is the rest.

640 Our main idea is finding the positions that can help discriminate the target family from the others
641 and assigning a high weight to these positions. Similarly, we assign a low weight to the positions
642 that are conserved for all subfamilies since they can cause incorrect hits. The details of how we
643 compute the weight per position is given in Algorithms 1 and 2.

644 In Step 1 of Algorithm 1, we show how to select the representative amino acids per group and
645 assign a score to each group based on the amino acid frequencies obtained from MSA for any
646 position k. For the target group, the most frequent amino acid, R_T, is chosen as the

representative one. For the closest and rest groups we first check if the group is composed of a single subfamily or multiple subfamilies. If the group is composed of a single subfamily, we select the most frequent amino acid as the representative and scores are taken as the frequency of this amino acid. Otherwise, we first check whether the most frequent amino acid for at least one element of the group is R_T. If it is, R_T is chosen as the representative amino acid for the corresponding group and its frequency is assigned as a group score. If not, the amino acid with the highest frequency for most of the elements is chosen, and the score is computed by taking the average of the frequencies of the representative amino acids in the subfamilies of the corresponding group.

Algorithm 1 – Step 2 shows the details of how we assign type, which represents whether the position can be used to discriminate the proteins in target and compute the initial score for any position k. To calculate the initial score, we defined six different categories and four different position types based on the representative amino acids that we identified in Step 1 of Algorithm 1. Type I, IV, III, II correspond to the positions ordered with respect to the highest to the lowest final weight.

- Category I (lines 12-18): Representative amino acids in the closest and rest groups are gaps. If the representative amino acid in the target group is also gap, then the initial weight type is Type II. Otherwise, the initial weight type is Type I since this position can be used to discriminate between target from close and rest groups.
- Category II (lines 19-25): The representative amino acid in the target group is gap. If the representative amino acid in the closest or rest group is not gap, then the initial weight type is Type I. Otherwise, the initial weight type is Type II.
- Category III (lines 26-56): Representative amino acids in the target, closest and rest groups are different from each other. We check boundary conditions to see whether each group is conserved. If all scores per group are greater than or equal to a

predefined threshold value, the position type is I. If any of them is smaller than the threshold, we check whether amino acid substitutions between the representative amino acids are probable with respect to the BLOSUM score. If the BLOSUM score between the representative amino acids of target and close or target and rest is less than the predefined threshold, we again assign Type I. Otherwise, the type is IV.

- Category IV (lines 57-59): The representative amino acids in the target and the representative amino acids in the closest groups are the same. The type is II. These positions are the main reason for the wrong hits, so we assign the lowest weights to the positions in this category.
- Category V (lines 60-70): The representative amino acid in the target group is the same with the representative amino acid in the rest group, but it is different from the representative amino acid in the closest group. If the conservation level of the rest is less than the threshold, the type is III, otherwise, to prevent wrong hits from the rest group to the target, we label the position as Type II.
- Category VI (lines 71-80): The representative amino acid in the closest group is the same with the representative amino acid in the rest group, but the representative amino acid in the target group is different. The type is I if it satisfies the boundary conditions; otherwise, the type is II.

The predefined thresholds thr_1 and thr_2 are chosen by considering the conservation dynamics of target and close&rest, respectively. For example, for CaSR, $thr_1 = 0.98$ and $thr_2 = 0.8$. For GPRC6A and TAS1Rs, we took both thr_1 and thr_2 as 0.8. The final threshold that is used to determine whether two amino acids are close to each other in terms of the BLOSUM score, thr_{bls}, is chosen as 0.5. Here, we do not directly use BLOSUM scores of two amino acids, instead we normalize each row of the log odd ratio of the BLOSUM matrix by dividing it by the maximum element of the corresponding row. Thus, the maximum value of the matrix is 1 when

697 the amino acid is conserved and the value decreases with respect to the closeness of
698 substituted amino acids.

699 After we check each category and assign types to the positions, we calculate the initial scores
700 that will be used to compute the final weight based on the position type. For Types I, III and IV,
701 the initial score is the sum of scores for each group. On the other hand, for Type II, the initial
702 score is computed as one over the sum of individual group scores. Here we aim to assign the
703 lowest weight to Type II since it consists of the positions that can cause wrong hits to target
704 because of similar conservation and amino acid patterns with close and/or rest groups.

705 After determining the types and initial score of each position, the next and final step is to decide
706 the final weight that will be used to modify the emission probabilities of the default HMM profiles.
707 Details of this process are given in Algorithm 2. The weight of each type will be in the following
708 order from highest to lowest: Type I, Type IV, Type III and Type II. We first order each position
709 in each group from highest to lowest with respect to the initial score. Type I positions take
710 weights greater than equal to c_1 where $c_1 = 1$ for CaSR, GPRC6A, TAS1R2 and TAS1R3 and
711 $c_1 = 1.5$ for TAS1R2. The maximum weight of Type IV positions, c_2, is equal to the mean value
712 of the weights of Type I positions (line 4 of Algorithm 2) for all target subfamilies. The maximum
713 weight of Type III positions, c_3, is taken as the minimum weight of Type IV positions for
714 GPRC6A and 0.5 for TAS1Rs. For CaSR, since it is a highly conserved family and Type I and IV
715 positions are much more important compared to the other types, we decreased the maximum
716 weight of Type III positions further by taking the minimum of Type IV over 2. Finally, Type II
717 positions take the lowest score as the highest weight, c_4, is again determined by considering the
718 conservation pattern of the target. We took $c_4 = 0.2$ for GPRC6A and TAS1R2, $c_4 = 0.25$ for
719 TAS1R1 and TAS1R3. For CaSR, the highest weight of Type II positions is restricted by the
720 minimum weight of Type III positions over 2. As mentioned earlier, this type includes positions

721 that show similar conservation patterns and cannot be used to distinguish between the elements

722 of the target and other groups.

ALGORITHM 1: REPRESENTATIVE AMINO ACID AND INITIAL SCORE FOR POSITION "K"

Input: *Representative amino acid of target subfamily, R_T; the frequency of R_T in the target, S_T; the most frequent amino acid of subfamily i, (i=1,...,N) and its frequency, a_i, F_i, respectively; target, close and rest groups, t, c, and r, respectively; the number of subfamilies in close and rest groups, n_c and n_r, respectively; conservation threshold for target and close/rest groups, thr_1 and thr_2; the threshold for Blosum scores, thr_{bls}.*

STEP 1: *Choose representative amino acid and related frequency for each group*

for $j \in \{c, r\}$
  if $n_j = 1$
    $R_j = a_k$ *where k is the subfamily in group j*
    $S_j = F_k$
  else
    if $R_T \in \{a_j, j = 1, \ldots, n_j\}$
      $R_j = R_T$
      $S_j = F_k$ *where k is the subfamily with the most frequent amino acid is R_T*
    else
      $R_j = a_k$ *where k is group with highest frequency*
      $S_j = \frac{\sum_{s=1}^{n_j} F_s}{n_j}$

STEP 2: *Assign position type and initial score to position "k"*

Category 1

if $R_c = R_r$ and they are gap
  if R_T is gap
    $type_k = II$
    $score_k = \frac{1}{\sum_{i \in \{T,c,r\}} S_i}$
  else
    $type_k = I$
    $score_k = \sum_{i \in \{T,c,r\}} S_i$

Category 2

else if R_T is gap
  if R_c is gap or R_r is gap
    $type_k = II$
    $score_k = \frac{1}{\sum_{i \in \{T,c,r\}} S_i}$

else

$type_k = I$

$score_k = \sum_{i \in \{T,c,r\}} S_i$

Category 3

else if $R_T \neq R_c \neq R_r$

if R_T, R_r *and* R_c *are not gaps*

if $S_T \geq thr_1$ *and* $S_c, S_r \geq thr_2$

$type_k = I$

$score_k = \sum_{i \in \{T,c,r\}} S_i$

else if $Blosum(R_T, R_c) \leq thr_{bls}$ *and* $Blosum(R_T, R_r) \leq thr_{bls}$

$type_k = I$

$score_k = \sum_{i \in \{T,c,r\}} S_i$

else

$type_k = IV$

$score_k = \sum_{i \in \{T,c,r\}} S_i$

else if R_c *is gap*

if $S_T \geq thr_1$ *and* $S_r \geq thr_2$

$type_k = I$

$score_k = \sum_{i \in \{T,c,r\}} S_i$

else if $Blosum(R_T, R_r) \leq thr_{bls}$

$type_k = I$

$score_k = \sum_{i \in \{T,c,r\}} S_i$

else

$type_k = IV$

$score_k = \sum_{i \in \{T,c,r\}} S_i$

else if R_r *is gap*

if $S_T \geq thr_1$ *and* $S_c \geq thr_2$

$type_k = I$

$score_k = \sum_{i \in \{T,c,r\}} S_i$

else if $Blosum(R_T, R_c) \leq thr_{bls}$

$type_k = I$

$score_k = \sum_{i \in \{T,c,r\}} S_i$

else

$type_k = IV$

$score_k = \sum_{i \in \{T,c,r\}} S_i$

Category 4

else if $R_T = R_c$

$type_k = II$

$score_k = \frac{1}{\sum_{i \in \{T,c,r\}} S_i}$

Category 5

else if $R_T \neq R_c$ *and* $R_T = R_r$

if R_c *is gap*

$type_k = III$

$score_k = \sum_{i \in \{T,c,r\}} S_i$

else

if $Blosum(R_T, R_c) \leq thr_{bls}$ *and* $S_c \geq thr_2$

$type_k = III$

$score_k = \sum_{i \in \{T,c,r\}} S_i$

else

$type_k = II$

$score_k = \frac{1}{\sum_{i \in \{T,c,r\}} S_i}$

Category 6

else if $R_T \neq R_c$ *and* $R_c = R_r$

if $S_T \geq thr_1$ *and* $S_c, S_r \geq thr_2$

$type_k = I$

$score_k = \sum_{i \in \{T,c,r\}} S_i$

else if $Blosum(R_T, R_c) \leq thr_{bls}$

$type_k = I$

$score_k = \sum_{i \in \{T,c,r\}} S_i$

else

$type_k = II$

$score_k = \frac{1}{\sum_{i \in \{T,c,r\}} S_i}$

ALGORITHM 2: COMPUTE WEIGHT FOR ALL POSITIONS OF TARGET SUBFAMILY "S"

Input: *Types for each position k (k=1,…,K),* $type_k$*; initial score for each position k of type t,* $score_{t,k}$*; number of type i positions,* n_i *where* $n_1 + n_2 + n_3 + n_4 = K$*; a predefined constant value as max weight of Type i positions,* c_i*; the target subfamily, S.*

Weight of Type I positions

for $p_1 = 1{:}\,n_1$

$weight_{p_1} = \frac{score_{1,p_1}}{\min_{l=1,\ldots,n_1}(score_{1,l})} c_1$

Weight of Type IV positions

for $p_2 = 1{:}\,n_2$

$c_2 = mean\left(weight_{p_1}\right)_{p_1 \in \{1,\ldots,n_1\}}$

$$weight_{p_2} = \frac{score_{2,p_2}}{\max\limits_{l=1,\ldots,n_2}(score_{2,l})} c_2$$

Weight of Type III positions

for $p_3 = 1{:}\,n_3$

$$weight_{p_3} = \frac{score_{3,p_3}}{\max\limits_{l=1,\ldots,n_3}(score_{3,l})} c_3$$

Weight of Type II positions

for $p_4 = 1{:}\,n_4$

$$weight_{p_4} = \frac{score_{4,p_4}}{\max\limits_{l=1,\ldots,n_4}(score_{4,l})} c_4$$

if S==GPRC6A

$$weight_{p_1} = wieght_{p_1} * 2$$

Subfamily Specific Position Scores

From the alignment we used to make subfamily specific profile HMMs, we randomly selected 264 CaSR like sequences (same number of sequences as CaSRs) and took all CaSR (264 proteins), GPRC6A (242 proteins) and TAS1Rs (TAS1R1 has 210, TAS1R2 has 173 and TAS1R3 has 273 proteins). We built an ML tree by using IQ-TREE multicore version 2.0.6 (Nguyen *et al*, 2015) with automatic model selection (Kalyaanamoorthy *et al*, 2017) (-m MFP) and ultrafast bootstrap (Hoang *et al*, 2018) (-bb 1000) parameters. For CaSR, GPRC6A, and TAS1Rs, we removed the positions from the MSA that correspond to a gap in the human receptor respectively. By using gap removed alignments and the ML tree, we did ancestral sequence reconstructions for each subfamily with IQ-TREE multicore version 2.0.6 with the -m JTT+R10 model parameter (Nguyen *et al.*, 2015). We showed specific residues that have a SDP score higher than 5 on the structures. We used the cyro-EM structure of CaSR (PDB:7DTV) and Swiss models (Waterhouse *et al*, 2018) for GPRC6A and taste receptors since they do not have experimental structures. To visualize structures and residues, we used the UCSF Chimera tool (Pettersen *et al*, 2004).

We calculated SDP scores by a method extended from (Bradley & Beltrao, 2019) by considering phylogenetic trees and a phylogeny-based scoring approach, adjPHACT, based on the

741 methodology of the PHACT algorithm. The details of how we compute the SDP score for any
742 position k can be found in Algorithm 3. PHACT computes the tolerance for each amino acid for
743 the query species, which is human, by using a tree traversal approach. By checking the
744 probability differences, PHACT detects the location of amino acid substitutions and compute the
745 weighted sum of positive probability differences based on the distance between the node of
746 change and human. On the other hand, here we aim to determine the acceptability of each
747 amino acid per subfamily. To achieve this, we modify PHACT by starting the tree traversal from
748 the root node and eliminating the node weighting approach. At the end, we have a probability
749 distribution per position for each subfamily, which is computed by considering the independent
750 events. Again, we determine the representative amino acid for the target subfamily by picking
751 the most frequently observed amino acid and its adjPHACT score. For the remaining
752 subfamilies, we keep the adjPHACT score of the representative amino acid of the target and the
753 representative amino acid of the corresponding subfamily. Then, similar to (Bradley & Beltrao,
754 2019) we check whether the same amino acid is conserved across all subfamilies. On the other
755 hand, our approach differs from (Bradley & Beltrao, 2019) in terms of considering multiple
756 subfamilies and using adjPHACT scores, which employ phylogenetic trees and ancestral
757 reconstruction probabilities. In our approach, we compute the contribution of each subfamily to
758 the SDP score by checking whether the representative amino acid of target has a high
759 adjPHACT score in that subfamily (line 2-7). In the final SDP score for any position k is
760 computed by considering the distance between target and other subfamilies (which is computed
761 by considering the distance between root nodes), the conservation level of the target subfamily
762 in terms of independent amino acid alterations and the individual score coming from each
763 subfamily (line 10).

ALGORITHM 3: SDP SCORE FOR POSITION "K"

Input: *Amino acid with the highest adjPHACT score in the target group,* a_T*; the adjPHACT score of* a_T *in the target,* $P_{a_T}^T$*; adjPHACT score of* a_T *in subfamily I (i=1,…,n),* $P_{a_T}^i$*; distance between target subfamily and subfamily I,* d_i*; amino acid with the highest adjPHACT score in the subfamily I,* a_i*; adjPHACT score of* a_i *in subfamily I,* $P_{a_i}^i$*.*

for i=1:n

Check whether the subfamily i is conserved

$$cons = -\left(P_{a_i}^i \leq 0.5\right) + \left(P_{a_i}^i > 0.5\right)$$

Specificity contribution of each subfamily $S_i = 0$

if $cons == 1$ *and* $a_i == a_T$

$$S_i = -\exp(P_{a_T}^i)$$

else if ($cons == -1$ *and* $a_i == a_T$*) or* $a_i \neq a_T$

$$S_i = \exp(1) - \exp(P_{a_i}^i)$$

end

The overall weight,

$$\omega = 1 - \max_{\mathrm{i=1,..,n}}\left(P_{a_T}^i\right).$$

The SDP score for position k,

$$SDP = \left(\exp\left(P_{a_T}^T\right) + \omega\left(\sum_{i=1}^{n}\frac{1}{D_i}S_i\right)\right)^{P_{a_T}^T}$$

Evolution of Class C GPCRs

We selected representative sequences from different taxonomic levels for each subfamily and 264 CaSR-like sequences. We aligned them with the MAFFT v7.221 einsi algorithm (Yamada *et al.*, 2016). We built the ML tree by RAxML-NG-0.9.0 with the model JTT and transfer bootstrap expectation –bs-metric fbp, tbe parameters (Kozlov *et al.*, 2019). We merged the ML trees of CaSR, GPRC6A and taste receptors by checking clades using the ETE toolkit (Huerta-Cepas *et al.*, 2016).

774 **Identification of the CaSR Activation Network**

775 To reveal the network that is important for CaSR activation, we measured changes in contact
776 scores between residues between active and inactive receptor states. For a single protomer, we
777 manipulated the pdb file and the RRCS algorithm (Zhou *et al.*, 2019), to be able to detect
778 changes in atom level for identical residues (python codes are provided). We applied a t-test to
779 identify significant changes and chose a p-value threshold of 0.01. Then we later combined
780 atomic-level changes observed within a protomer for each residue to build the activation
781 network. If atoms of a residue are involved in multiple significant changes, this is represented as
782 multiple edges in the network, even if the residue pair is the same.

783 For the analysis of interactions observed inside the dimerization interface, we used the RRCS
784 algorithm as is and compared residue-residue contact scores between residue pairs upon
785 activation. The important thing to note here is that for every dimer structure we used, we
786 sometimes retrieved two data points due to the symmetrical nature of the dimerization. We
787 again applied the same p-value threshold and identified the important changes observed.

788 In both of our analyses, we used same set of 7 active state (PDB IDs 7SIL(Park *et al*, 2021b),
789 7SIM(Park *et al.*, 2021b), 7E6T(Chen *et al.*, 2021), 7M3G(Gao *et al.*, 2021), 7M3F(Gao *et al.*,
790 2021), 7DTT(Ling *et al.*, 2021), 7DTV(Ling *et al.*, 2021)) and 5 inactive state structures
791 (7SIN(Park *et al.*, 2021a), 7E6U(Chen *et al.*, 2021), 7M3E(Gao *et al.*, 2021), 7M3J(Gao *et al.*,
792 2021), 7DTW(Ling *et al.*, 2021)) human CaSR structures.

793 **Machine Learning**

794 **Dataset and Feature Preparation**

795 To predict the consequence of a substitution in human CaSR, we used a gradient boosting-
796 based machine learning algorithm, XGBoost (Chen & Guestrin, 2016). We used the XGBoost
797 library for R (Chen *et al*, 2015) to train our model. We selected a total of 337 LoF and GoF
798 mutations from the literature (Gorvin, 2019) to train our model. Since we used conservation

799 scores as features to train our model, we divided subfamily alignments and mutations randomly
800 as 80% training and the remaining 20% test data before creating feature matrices to prevent
801 information leakage. 25% of the training data was randomly picked as the validation data five
802 times for cross-validation. For each dataset split we used the sklearn train test split model with
803 stratify option to keep the LoF to GoF ratio almost the same in the datasets (Pedregosa *et al*,
804 2011). We calculated the conservation score of the reference amino acid and the substituted
805 amino acid in human CaSR in each subfamily. The reference and the substituted amino acids
806 were represented by BLOSUM62-encoded matrices. Amino acid physico-chemical feature
807 values Zimmerman polarity (Zimmerman *et al*, 1968), average flexibility (Bhaskaran &
808 Ponnuswamy, 1988), Dayhoff (Barker *et al*, 1978), average buried area (Rose *et al*, 1985),
809 Doolittle hydropathicity (Kyte & Doolittle, 1982), atomic weight ratio (Grantham, 1974),
810 molecular weight, and bulkiness (Zimmerman *et al.*, 1968) from the ProtScale database
811 (Walker, 2005); and domain information of the reference amino acid were used as other
812 features. We normalized the physico-chemical feature values prior to model training. We
813 repeated the whole random dataset splitting and feature preparation procedure 50 times to
814 obtain more robust results.

815

816 **Model Selection and Parameter Tuning**

817 We picked the model parameters for each replication by applying a 5-fold cross-validation
818 technique to the training set. We tuned the model parameters step-by-step using the same
819 validation sets for each parameter to decrease the time complexity. We used the following order
820 of model parameters, so that the parameter that has the highest impact on model outcome was
821 tuned first: Eta and nrounds, gamma, maxdepth, subsample, colsample bytree, min child
822 weight, lambda, alpha. We selected the maxdepth as 2, the minimum maxdepth value to
823 prevent overfitting. We chose eta, gamma, colsample bytree, subsample, min child weight from

824 the sets 0.00001,0.00002,..., 0.001,0,0.1,0.2,...,0.5, 0.5,0.55,...,1, 0.5,0.55,...,1, 1,2,...,6
825 respectively. We selected regularization parameters lambda and alpha from the set 0, 1e-4, 1e-
826 3, 1e-2, 1e-1, 1, 10, 100. We set the nrounds parameter to 200.
827

828 **Performance Metrics**

829 We used the area under the receiver operating characteristic curve (AUROC) and the area
830 under the precision-recall curve (AUPR) to evaluate the performance of our prediction model.
831 AUROC and AUPR are performance measures that are widely used to evaluate the
832 performance of binary classification problems. The higher the AUROC and AUPR, the better the
833 model distinguishes classes. To understand how our model makes predictions, we used SHAP
834 (SHapley Additive exPlanations) values. Shap values give an estimate of how much ach feature
835 contributed to the prediction of the model made (Lundberg & Lee, 2017). We calculated SHAP
836 values for our final model trained by all samples by the using R shapviz package (Mayer, 2023).
837

838 **Predictive Performance**

839 After we evaluated the performance of our machine learning algorithm over 50 replications, we
840 used the whole dataset to train the model that we used to make predictions for every possible
841 mutation in human CaSR. We selected model parameters by using the 5-fold cross-validation
842 technique on the whole dataset. To create a new test dataset, we took subfamily alignments of
843 the species from the new Uniprot dataset that did not exist in the training data. We eliminated
844 amino acids that are observed in the CaSR alignment as neutral. In each position, we predicted
845 the GoF or LoF class for any substitution. We did a literature search to find new clinical cases
846 that cause either GoF or LoF mutations. We reported our predictions in the table.
847

848 **Data Availability**

849 The open-source code is available at our GitHub repository:

850 https://github.com/CompGenomeLab/CaSR

851

852 **Acknowledgments**

853 This study was supported by EMBO Installation Grant no:4163 funded by TÜBİTAK (to OA).

854 Ogun Adebali is supported by the BAGEP program of the Science Academy - Türkiye, and the

855 TÜBA-GEBİP program of the Turkish Academy of Sciences.

856

857 **References**

858 Alirezaie N, Kernohan KD, Hartley T, Majewski J, Hocking TD (2018) ClinPred: prediction tool to
859 identify disease-relevant nonsynonymous single-nucleotide variants. *The American Journal of*
860 *Human Genetics* 103: 474-483
861 Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool.
862 *Journal of molecular biology* 215: 403-410
863 Barker WC, Ketcham LK, Dayhoff MO (1978) A comprehensive examination of protein
864 sequences for evidence of internal gene duplication. *Journal of Molecular Evolution* 10: 265-281
865 Bhaskaran R, Ponnuswamy P (1988) Positional flexibilities of amino acid residues in globular
866 proteins. *International Journal of Peptide and Protein Research* 32: 241-255
867 Bletsis P, Metzger R, Nelson JA, Gasparini J, Alsayed M, Milas M (2022) A Novel missense
868 CASR gene sequence variation resulting in familial hypocalciuric hypercalcemia. *AACE Clinical*
869 *Case Reports* 8: 194-198
870 Bradley D, Beltrao P (2019) Evolution of protein kinase substrate recognition at the active site.
871 *PLoS biology* 17: e3000341
872 Brown D, Krishnamurthy N, Dale JM, Christopher W, Sjolander K (2005) Subfamily hmms in
873 functional genomics. *Pac Symp Biocomput*: 322-333
874 Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T (2009) trimAl: a tool for automated
875 alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25: 1972-1973
876 Chagoyen M, Garcia-Martin JA, Pazos F (2016) Practical analysis of specificity-determining
877 residues in protein families. *Brief Bioinform* 17: 255-261

878 Chen T, Guestrin C, 2016. Xgboost: A scalable tree boosting system, Proceedings of the 22nd
879 acm sigkdd international conference on knowledge discovery and data mining. pp. 785-794.
880 Chen T, He T, Benesty M, Khotilovich V, Tang Y, Cho H, Chen K (2015) Xgboost: extreme
881 gradient boosting. R package version 04-2. *OS Independent*
882 Chen X, Wang L, Cui Q, Ding Z, Han L, Kou Y, Zhang W, Wang H, Jia X, Dai M *et al* (2021)
883 Structural insights into the activation of human calcium-sensing receptor. *Elife* 10
884 Chou KJ, Hsu CY, Huang CW, Chen HJ, Ou SH, Chen CL, Lee PT, Fang HC (2022) A new
885 missense mutation of calcium sensing receptor with isoleucine replaced by serine at codon 857
886 leading to type V Bartter syndrome. *Exp Cell Res* 414: 113080
887 Chun L, Zhang WH, Liu JF (2012) Structure and ligand recognition of class C GPCRs. *Acta*
888 *Pharmacol Sin* 33: 312-323
889 Consortium TU (2017) UniProt: the universal protein knowledgebase. *Nucleic acids research*
890 45: D158-D169
891 Cook AE, Mistry SN, Gregory KJ, Furness SGB, Sexton PM, Scammells PJ, Conigrave A,
892 Christopoulos A, Leach K (2015) Biased allosteric modulation at the CaS receptor engendered
893 by structurally diverse calcimimetics. *British journal of pharmacology* 172: 185-200
894 Coughlan A, Khan F, Brassill M (2022) A Novel Genetic Variant Resulting in Familial
895 Hypocalciuric Hypercalcaemia. *Irish Medical Journal* 115: 545-545
896 Courtney A, Hill A, Smith D, Agha A (2022) Familial hypocalciuric hypercalcaemia type 1 caused
897 by a novel heterozygous missense variant in the CaSR gene, p (His41Arg): two case reports.
898 *BMC Endocrine Disorders* 22: 324
899 Dharmaraj P, Gorvin CM, Soni A, Nelhans ND, Olesen MK, Boon H, Cranston T, Thakker RV,
900 Hannan FM (2020) Neonatal Hypocalcemic Seizures in Offspring of a Mother With Familial
901 Hypocalciuric Hypercalcemia Type 1 (FHH1). *J Clin Endocrinol Metab* 105
902 Dong Q, Song F, Du M, Qiu M, Chen X (2020) [Clinical and genetic analysis of a child with
903 neonatal severe parathyroidism]. *Zhonghua Yi Xue Yi Chuan Xue Za Zhi* 37: 1247-1249
904 Eddy SR (2011) Accelerated profile HMM searches. *PLoS computational biology* 7: e1002195
905 Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, Potter SC, Punta M, Qureshi
906 M, Sangrador-Vegas A (2016) The Pfam protein families database: towards a more sustainable
907 future. *Nucleic acids research* 44: D279-D285
908 Flock T, Hauser AS, Lund N, Gloriam DE, Balaji S, Babu MM (2017) Selectivity determinants of
909 GPCR-G-protein binding. *Nature* 545: 317-322

910 Gao Y, Robertson MJ, Rahman SN, Seven AB, Zhang C, Meyerowitz JG, Panova O, Hannan
911 FM, Thakker RV, Brauner-Osborne H *et al* (2021) Asymmetric activation of the calcium-sensing
912 receptor homodimer. *Nature* 595: 455-459
913 Geng Y, Mosyak L, Kurinov I, Zuo H, Sturchler E, Cheng TC, Subramanyam P, Brown AP,
914 Brennan SC, Mun HC *et al* (2016) Structural mechanism of ligand activation in human calcium-
915 sensing receptor. *Elife* 5
916 Goes van Naters W, Mucignat-Caretta C (2014) Frontiers in Neuroscience Drosophila
917 Pheromones: From Reception to Perception. *Neurobiology of Chemical Communication Boca*
918 *Raton (FL): CRC Press/Taylor & Francis (c)*
919 Gomes V, Silvestre C, Ferreira F, Bugalho M (2020) Autosomal dominant hypocalcaemia:
920 identification of two novel variants of CASR gene. *BMJ Case Rep* 13
921 Gorvin CM (2019) Molecular and clinical insights from studies of calcium-sensing receptor
922 mutations. *J Mol Endocrinol* 63: R1-R16
923 Gorvin CM (2021) Calcium-sensing receptor signaling—How human disease informs biology.
924 *Current opinion in endocrine and metabolic research* 16: 10-18
925 Grantham R (1974) Amino acid difference formula to help explain protein evolution. *science*
926 185: 862-864
927 Hao Y, Lei Z, Shi N, Yu L, Ji W, Zhang X (2021) Radiofrequency Ablation of Parathyroid Glands
928 to Treat a Patient With Hypercalcemia Caused by a Novel Inactivating Mutation in CaSR. *Front*
929 *Endocrinol (Lausanne)* 12: 743517
930 Harpsoe K, Boesgaard MW, Munk C, Brauner-Osborne H, Gloriam DE (2017) Structural insight
931 to mutation effects uncover a common allosteric site in class C GPCRs. *Bioinformatics* 33: 1116-
932 1120
933 Hauser AS, Kooistra AJ, Munk C, Heydenreich FM, Veprintsev DB, Bouvier M, Babu MM,
934 Gloriam DE (2021) GPCR activation mechanisms across classes and macro/microscales. *Nat*
935 *Struct Mol Biol* 28: 879-888
936 Hoang DT, Chernomor O, Von Haeseler A, Minh BQ, Vinh LS (2018) UFBoot2: improving the
937 ultrafast bootstrap approximation. *Molecular biology and evolution* 35: 518-522
938 Huerta-Cepas J, Serra F, Bork P (2016) ETE 3: reconstruction, analysis, and visualization of
939 phylogenomic data. *Molecular biology and evolution* 33: 1635-1638
940 Ioannidis NM, Rothstein JH, Pejaver V, Middha S, McDonnell SK, Baheti S, Musolf A, Li Q,
941 Holzinger E, Karyadi D (2016) REVEL: an ensemble method for predicting the pathogenicity of
942 rare missense variants. *The American Journal of Human Genetics* 99: 877-885

943 Jones EM, Lubock NB, Venkatakrishnan A, Wang J, Tseng AM, Paggi JM, Latorraca NR,
944 Cancilla D, Satyadi M, Davis JE (2020) Structural and functional characterization of G protein–
945 coupled receptors with deep mutational scanning. *Elife* 9: e54895
946 Kalyaanamoorthy S, Minh BQ, Wong TK, Von Haeseler A, Jermiin LS (2017) ModelFinder: fast
947 model selection for accurate phylogenetic estimates. *Nature methods* 14: 587-589
948 Kozlov AM, Darriba D, Flouri T, Morel B, Stamatakis A (2019) RAxML-NG: a fast, scalable and
949 user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics* 35: 4453-4455
950 Kumari R, Castillo C, Francesconi A (2013) Agonist-dependent signaling by group I
951 metabotropic glutamate receptors is regulated by association with lipid domains. *J Biol Chem*
952 288: 32004-32019
953 Kuru N, Dereli O, Akkoyun E, Bircan A, Tastan O, Adebali O (2022) PHACT: Phylogeny-Aware
954 Computing of Tolerance for Missense Mutations. *Mol Biol Evol* 39
955 Kyte J, Doolittle RF (1982) A simple method for displaying the hydropathic character of a
956 protein. *Journal of molecular biology* 157: 105-132
957 Li N, Li X, Ni XL, Li XY, Xia WB, Yang GQ, Pei Y (2021) A novel homozygous mutation of the
958 calcium-sensing receptor gene associated with apparent autosomal recessive inheritance of
959 familial hypocalciuric hypercalcemia. *Chin Med J (Engl)* 134: 1869-1871
960 Ling S, Shi P, Liu S, Meng X, Zhou Y, Sun W, Chang S, Zhang X, Zhang L, Shi C *et al* (2021)
961 Structural mechanism of cooperative activation of the human calcium-sensing receptor by
962 Ca(2+) ions and L-tryptophan. *Cell Res* 31: 383-394
963 Liu H, Yi P, Zhao W, Wu Y, Acher F, Pin J-P, Liu J, Rondard P (2020) Illuminating the allosteric
964 modulation of the calcium-sensing receptor. *Proceedings of the National Academy of Sciences*
965 117: 21711-21722
966 Lundberg SM, Lee S-I (2017) A unified approach to interpreting model predictions. *Advances in*
967 *neural information processing systems* 30
968 Magno AL, Leatherbarrow KM, Brown SJ, Wilson SG, Walsh JP, Ward BK (2020) Functional
969 Analysis of Calcium-Sensing Receptor Variants Identified in Families Provisionally Diagnosed
970 with Familial Hypocalciuric Hypercalcaemia. *Calcif Tissue Int* 107: 230-239
971 Majumdar SK, Jacob T, Bale A, Bailey A, Kwon J, Hughes T, Barbieri AL, Laskin W, Cohen P,
972 Carling TJE (2020) A Novel Variant in the Calcium-Sensing Receptor Associated with Familial
973 Hypocalciuric Hypercalcemia and Low-to-Normal PTH. *Case Rep Endocrinol* 2020: 8752610
974 Mayer M, 2023. shapviz: SHAP Visualizations. R package version 0.9.0.

975 Moon JE, Yang HY, Wee G, Par KS, Ko CW (2021) A cell function study on calcium regulation of
976 a novel calcium-sensing receptor mutation (p.Tyr825Phe). *Ann Pediatr Endocrinol Metab* 26:
977 24-30
978 Mullin BH, Pavlos NJ, Brown SJ, Walsh JP, McKellar RA, Wilson SG, Ward BK (2022)
979 Functional assessment of calcium-sensing receptor variants confirms familial hypocalciuric
980 hypercalcemia. *Journal of the Endocrine Society* 6: bvac025
981 Mun H-C, Culverston EL, Franks AH, Collyer CA, Clifton-Bligh RJ, Conigrave AD (2005) A
982 double mutation in the extracellular Ca2+-sensing receptor's venus flytrap domain that
983 selectively disables L-amino acid sensing. *Journal of Biological Chemistry* 280: 29067-29072
984 Nango E, Akiyama S, Maki-Yonekura S, Ashikawa Y, Kusakabe Y, Krayukhina E, Maruno T,
985 Uchiyama S, Nuemket N, Yonekura K *et al* (2016) Taste substance binding elicits
986 conformational change of taste receptor T1r heterodimer extracellular domains. *Sci Rep* 6:
987 25745
988 Nguyen L-T, Schmidt HA, Von Haeseler A, Minh BQ (2015) IQ-TREE: a fast and effective
989 stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular biology and*
990 *evolution* 32: 268-274
991 Nuemket N, Yasui N, Kusakabe Y, Nomura Y, Atsumi N, Akiyama S, Nango E, Kato Y, Kaneko
992 MK, Takagi J *et al* (2017) Structural basis for perception of diverse chemical substances by T1r
993 taste receptors. *Nat Commun* 8: 15530
994 Palmieri S, Grassi G, Guarnieri V, Chiodini I, Arosio M, Eller-Vainicher C (2021) Case Report:
995 Unusual Presentations of Loss-of-Function Mutations of the Calcium-Sensing Receptor. *Front*
996 *Med (Lausanne)* 8: 809067
997 Park J, Zuo H, Frangaj A, Fu Z, Yen LY, Zhang Z, Mosyak L, Slavkovich VN, Liu J, Ray KM *et al*
998 (2021a) Symmetric activation and modulation of the human calcium-sensing receptor. *Proc Natl*
999 *Acad Sci U S A* 118
1000 Park J, Zuo H, Frangaj A, Fu Z, Yen LY, Zhang Z, Mosyak L, Slavkovich VN, Liu J, Ray KM *et al*
1001 (2021b) Symmetric activation and modulation of the human calcium-sensing receptor.
1002 *Proceedings of the National Academy of Sciences* 118: e2115849118
1003 Patil I (2021) Visualizations with statistical details: The'ggstatsplot'approach. *Journal of Open*
1004 *Source Software* 6: 3167
1005 Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer
1006 P, Weiss R, Dubourg V (2011) Scikit-learn: Machine learning in Python. *the Journal of machine*
1007 *Learning research* 12: 2825-2830

1008 Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE (2004)
1009 UCSF Chimera--a visualization system for exploratory research and analysis. *J Comput Chem*
1010 25: 1605-1612
1011 Pi M, Nishimoto SK, Quarles LD (2017) GPRC6A: Jack of all metabolism (or master of none).
1012 *Mol Metab* 6: 185-193
1013 Pin JP, Galvez T, Prezeau L (2003) Evolution, structure, and activation mechanism of family 3/C
1014 G-protein-coupled receptors. *Pharmacol Ther* 98: 325-354
1015 Pruitt KD, Tatusova T, Maglott DR (2007) NCBI reference sequences (RefSeq): a curated non-
1016 redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* 35:
1017 D61-65
1018 Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M (2019) CADD: predicting the
1019 deleteriousness of variants throughout the human genome. *Nucleic acids research* 47: D886-
1020 D894
1021 Roberts MS, Gafni RI, Brillante B, Guthrie LC, Streit J, Gash D, Gelb J, Krusinska E, Brennan
1022 SC, Schepelmann M *et al* (2019) Treatment of Autosomal Dominant Hypocalcemia Type 1 With
1023 the Calcilytic NPSP795 (SHP635). *J Bone Miner Res* 34: 1609-1618
1024 Rogers MF, Shihab HA, Mort M, Cooper DN, Gaunt TR, Campbell C (2018) FATHMM-XF:
1025 accurate prediction of pathogenic point mutations via extended features. *Bioinformatics* 34: 511-
1026 513
1027 Rose GD, Geselowitz AR, Lesser GJ, Lee RH, Zehfus MH (1985) Hydrophobicity of amino acid
1028 residues in globular proteins. *Science* 229: 834-838
1029 Sagi SV, Joshi H, Trotman J, Elsey T, Swamy A, Rajkanna J, Bhat NA, Haddadin FJS, Oyibo
1030 SO, Park SM (2020) A novel CASR variant in a family with familial hypocalciuric hypercalcaemia
1031 and primary hyperparathyroidism. *Endocrinol Diabetes Metab Case Rep* 2020
1032 Srivastava PK, Desai DK, Nandi S, Lynn AM (2007) HMM-ModE--improved classification using
1033 profile hidden Markov models by optimising the discrimination threshold and modifying emission
1034 probabilities with negative training sequences. *BMC Bioinformatics* 8: 104
1035 Stamatakis A (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of
1036 large phylogenies. *Bioinformatics* 30: 1312-1313
1037 Studer RA, Dessailly BH, Orengo CA (2013) Residue mutations and their impact on protein
1038 structure and function: detecting beneficial and pathogenic changes. *Biochemical journal* 449:
1039 581-594

1040 Tsuji T, Hiroyuki A, Uraki S, Doi A, Morita S, Iwakura H, Nishi M, Furuta H, Akamizu T (2021)
1041 Autosomal Dominant Hypocalcemia With Atypical Urine Findings Accompanied by Novel CaSR
1042 Gene Mutation and VitD Deficiency. *J Endocr Soc* 5: bvaa190
1043 Walker JM (2005) *The proteomics protocols handbook*. Springer
1044 Waterhouse A, Bertoni M, Bienert S, Studer G, Tauriello G, Gumienny R, Heer FT, de Beer TAP,
1045 Rempfer C, Bordoli L (2018) SWISS-MODEL: homology modelling of protein structures and
1046 complexes. *Nucleic acids research* 46: W296-W303
1047 Wejaphikul K, Dejkhamron P, Khorana J, Watcharachan K, Intachai W, Olsen B, Tongsima S,
1048 Cairns JRK, Ngamphiw C, Kantaputra P (2023) Subtotal parathyroidectomy successfully
1049 controls calcium levels of patients with neonatal severe hyperparathyroidism carrying a novel
1050 CASR mutation. *Hormone Research in Paediatrics*: 1-7
1051 Wen T, Wang Z, Chen X, Ren Y, Lu X, Xing Y, Lu J, Chang S, Zhang X, Shen Y (2021)
1052 Structural basis for activation and allosteric modulation of full-length calcium-sensing receptor.
1053 *Science Advances* 7: eabg1483
1054 Wootten D, Christopoulos A, Marti-Solano M, Babu MM, Sexton PM (2018) Mechanisms of
1055 signalling and biased agonism in G protein-coupled receptors. *Nat Rev Mol Cell Biol* 19: 638-
1056 653
1057 Wu Y, Zhang C, Huang X, Cao L, Liu S, Zhong P (2022) Autosomal dominant hypocalcemia with
1058 a novel CASR mutation: a case study and literature review. *Journal of International Medical*
1059 *Research* 50: 03000605221110489
1060 Yamada KD, Tomii K, Katoh K (2016) Application of the MAFFT sequence alignment program to
1061 large data—reexamination of the usefulness of chained guide trees. *Bioinformatics* 32: 3246-
1062 3251
1063 Zajickova K, Dvorakova M, Moravcova J, Vcelak J, Goltzman D (2020) Familial hypocalciuric
1064 hypercalcemia in an index male: grey zones of the differential diagnosis from primary
1065 hyperparathyroidism in a 13-year clinical follow up. *Physiol Res* 69: S321-S328
1066 Zhang C, Huang Y, Jiang Y, Mulpuri N, Wei L, Hamelberg D, Brown EM, Yang JJ (2014)
1067 Identification of an L-phenylalanine binding site enhancing the cooperative responses of the
1068 calcium-sensing receptor to calcium. *Journal of Biological Chemistry* 289: 5296-5309
1069 Zhang C, Zhang T, Zou J, Miller CL, Gorkhali R, Yang JY, Schilmiller A, Wang S, Huang K,
1070 Brown EM *et al* (2016) Structural basis for regulation of human calcium-sensing receptor by
1071 magnesium ions and an unexpected tryptophan derivative co-agonist. *Sci Adv* 2: e1600241

1072 Zhang Z, Qiu W, Quinn SJ, Conigrave AD, Brown EM, Bai M (2002) Three adjacent serines in
1073 the extracellular domains of the CaR are required for L-amino acid-mediated potentiation of
1074 receptor function. *Journal of Biological Chemistry* 277: 33727-33735
1075 Zhou Q, Yang D, Wu M, Guo Y, Guo W, Zhong L, Cai X, Dai A, Jang W, Shakhnovich EI *et al*
1076 (2019) Common activation mechanism of class A GPCRs. *Elife* 8
1077 Zimmerman J, Eliezer N, Simha R (1968) The characterization of amino acid sequences in
1078 proteins by statistical methods. *Journal of theoretical biology* 21: 170-201

1079