Evolutionary history of Calcium-sensing receptors sheds light into hyper/hypocalcemia-causing mutations

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

1 Faculty of Engineering and Natural Sciences, Sabanci University, İstanbul, Türkiye
2 TÜBİTAK Research Institute for Fundamental Sciences, Gebze, Türkiye

* oadebali@sabanciuniv.edu

Abstract

The Calcium Sensing Receptor (CaSR) is very important in controlling the levels of calcium in the body by interacting with different types of G-protein. This receptor is highly conserved among other G-protein coupled receptors (GPCRs) and has been linked to disorders affecting the balance of calcium in the body, such as hypercalcemia and hypocalcemia. Although there has been progress in understanding the structure and function of CaSR, there is still a lack of knowledge about which specific residues are important for their function and how it differs from other receptors in the same class. In this study, we used phylogeny-based methods to identify functionally-equivalent orthologs of CaSR, predict the importance of each residue, and calculate specificity-determining position (SDP) scores to uncover the evolutionary basis of its function. Our results showed that the CaSR subfamily is highly conserved, with higher SDP scores than its closest receptor subfamilies. Residues with high SDP scores are likely to be critical in receptor activation and pathogenicity. We applied gradient-boosting trees with evolutionary metrics as inputs to predict the functional consequences of each substitution, and discriminate between gain and loss-of-function mutations those causing hypo- and hypercalcemia, respectively. Our study provides insight into the evolutionary fine-tuning of CaSR, which can help understand its role in calcium balance and related disorders.

Introduction

Calcium sensing receptor (CaSR) is a class C G-protein-coupled receptor (GPCR) that maintains extracellular Ca2+ homeostasis by sensing calcium ions in the blood and regulating parathyroid hormone release and urinary calcium[1, 2] . The CaSR is activated by Ca2+ and L-amino acids such as L-Phe and L-Trp as well as polyamines and polypeptides[3-5]. Ligands bind to the extracellular Venus flytrap (VFT) domain of the receptor-like the other class C GPCRs such as metabotropic glutamate receptors [6].

Class C GPCRs are obligatory dimers, forming either homo or heterodimers [6]. CaSR forms a homodimer where each subunit is composed of an extracellular domain (ECD), comprising a bilobed (LB1, LB2) VFT and a cysteine-rich domain (CRD) connected to a heptahelical transmembrane (7TM) domain[3, 5].

Crystal structures of the ECD [4, 7] and cryo-electron microscopy structures of the full-length CaSR[3, 5, 8-10] reveal the structural basis for activation mechanisms and ligand binding sites. L-amino acid binding site at the interdomain cleft of LB1-LB2[3-5, 11-13] and multiple Ca2+ amino acid binding sites on the VFT domain are shown in the literature[3-5]. While Ca2+ is the composite agonist to the CaSR, L-amino acids promote the receptor activation along with the Ca2+, but they are not able to activate the receptor alone[14]. Even though Ca2+ alone activates the receptor in functional assays[14], whether it activates the CaSR in the absence of L-amino acid is still controversial[3, 5].

Variants in CaSR may cause malfunctions that result in Ca2+ homeostasis diseases in humans. More than 400 germline loss/gain-of-function mutations cause hypercalcaemic disorders, neonatal severe hyperparathyroidism (NSHPT) and the milder familial hypocalciuric hypercalcemia type-1 (FHH1) and autosomal dominant hypocalcemia type-1 (ADH1) respectively[2]. Many more CaSR variants are anticipated to be identified as more population-level genetic data become available[2]. Understanding the role of each residue in receptor structure and activation mechanisms could provide additional information about the likelihood of variant pathogenicity and CaSR signaling. The role of each residue in a receptor can be revealed by comparison of receptors in a family and between different families; however, the structure and complete activation mechanisms of many families in class C GPCRs are still unknown, especially G-protein coupled receptor family C group 6 member A (GPRC6A) and type 1 taste receptors (TAS1Rs; members 1,2 and 3) that are the closest subfamilies to CaSR.

While all subfamily receptors of class C GPCRs share common domains and structural features, details of responding to different ligands and activating signaling pathways are diverse between even closely related receptors[6]. Gene duplication is the main mechanism that generates new protein functions across GPCRs. Protein families are evolved by speciation events following a gene duplication[15, 16]; thus sequence comparisons of members within a subfamily and between subfamilies can show the evolutionarily conserved domains as well as diverged protein sites that distinguish one subfamily from others. One challenge with this analysis is that excessive gene duplication events complicate the identification of functionally identical orthologs in a subfamily. Moreover, the conservation patterns in paralogs and distant homologs may help

inferring the specific roles of a single residue in protein function. Because the evolutionary pressure on the paralogs and close orthologues are not the same, allowed substitutions on paralogs may not be acceptable in close orthologues. Thus, using functionally identical orthologs in sequence comparisons is crucial to infer the role of each residue in a protein family.

Here, we show the importance of each residue in CaSR by comparing it with the closely related subfamilies, GPRC6A and TAS1Rs. We identified all orthologues sequences in each subfamily by phylogenetic tree analysis. To obtain orthologues without requiring computationally expensive phylogenetic tree step, subfamily-specific profile HMMs are generated from the true orthologues in subfamilies that we determined by the phylogenetic tree analysis. We calculated a specificity score for each residue in a subfamily by calculating scores based on a modified version of PHACT[17] scores which considers independent evolutionary events on the phylogenetic tree while scoring the acceptability of an amino acid substitution. We predicted the functional consequence of each substitution in CaSR by using the gradient boosting trees machine learning approach.

Results

Evolutionary History of Class C GPCRs

To reveal the evolutionary constraints on protein families, we developed to develop a strategy to precisely define a protein subfamily. Precise subfamily definition can be precisely accomplished by revealing the evolutionary history of the superfamily. Evolutionary history of gene families can only be established by reconstructing high-quality phylogenetic trees, which can be used to pinpoint gene duplication events. Discrimination between gene duplication and speciation nodes enabled us to define of the paralogous and orthologous protein sequences. We further analyzed the phylogenetic trees to classify the orthologous sequences that are likely equivalent in function. We used functionally-equivalent orthologs in comparative analyses between subfamilies, which eventually yielded subfamily-specific signatures that can be used to define that particular subfamily and its function. Finally, the association between the signature and function would enable a better understanding of specific molecular mechanisms and the effects of variants, particularly for the protein subfamily of interest. Here, we aim to reveal the signatures of the CaSR subfamily, that is implied in the specific function of calcium-sensing and downstream signaling.

We have retrieved the complete proteomes of 478 species from the NCBI database. To identify proteins that belong to class C GPCR family, we performed searched profile HMM of the seven transmembrane domain profile (Pfam: 7tm_3) (Fig 1) against the proteomes. While this search

allowed us to retrieve the entire class C GPCRs hitting the hmm profile, it does not yield subfamily (22 human GPCRs) annotations. We performed scan profile HMM of the PfamA profile against to Class C GPCR to select canonical isoforms. We used seven-transmembrane domain only to assign subfamilies and trimmed the N-terminus region of this domain for further homology steps. To generate a general HMM profile for each subfamily, we first applied a Blast search using each human class-C as a query. For each subject, we blasted them against the human proteome. We retrieved the bidirectional best hits (Core subfamily assignment).

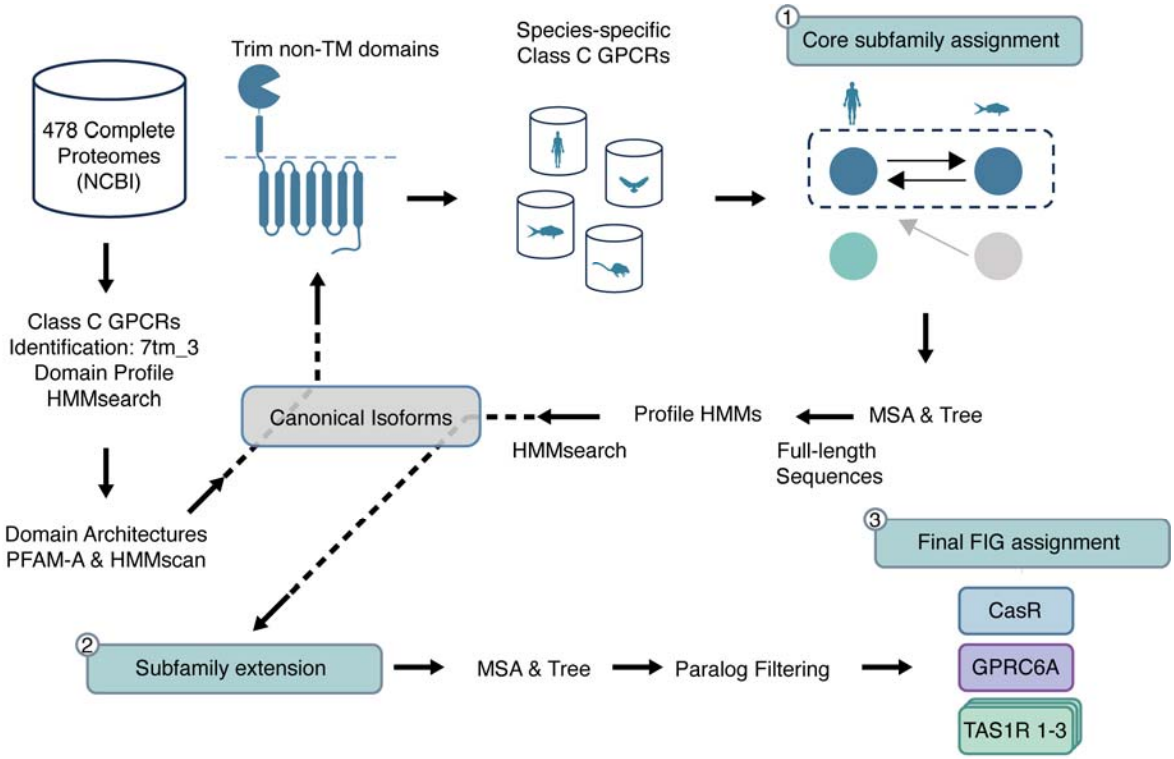

Figure 1: ***Summary of the Methodological Framework****. 478 complete proteomes were retrieved from NCBI database. Each sequence was searched by HMMsearch against Pfam 7tm_3 domain profile to retrieve all class C GPCRs. Domain architectures of class C GPCRs were determined by HMMscan against PFAM.A profile to identify canonical isoforms. Species-specific BLAST databases from TM domains 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). 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). We produced maximum likelihood (ML) trees of extended subfamilies and filtered paralogous sequences to obtain functionally identical groups (FIGs).

The CaSR subfamily produced over five thousand hits, which included vomeronasal and olfactory receptors that have never been shown to sense calcium. Previous research has shown that CaSR is classified in the pheromone/olfactory cluster of class C GPCRs[18] (18). In species that had multiple proteins assigned to the CaSR subfamily, we constructed a maximum likelihood tree using these hits and other human class C GPCR protein sequences. These trees revealed that a significant number of duplication events occurred in the species after the clade diverged from CaSR. As a result, we defined this diverged clade as a new subfamily named CaSR-likes. The sequences in this subfamily is unlikely to maintain calcium homeostasis, and therefore should not be annotated as calcium-sensing receptors.

We selected representative sequences from different species for each subfamily of 22 different receptor subfamilies and 264 CaSR-like sequences and built a ML tree (Fig 1A). Also, we built the ML trees of all proteins from CaSR, GPRC6A, taste receptors and merged these trees to the representative tree of class C GPCRs (Fig 2 A). The resulting phylogeny shows that are five major clades: CaSR-related, GABA, mGluR, Orphans, and retinoic acid induced (RAIG). Orphan receptors, GPR158 and GPR179, formed a clade that was diverged from other receptors consistent with previous trees[19] and with 0.95 transfer bootstrap (TB) value. γ-aminobutyric acidB receptors (GABBR1 and GABBR2) formed another clade diverged from GPR156 with 0.97 TB. γ-aminobutyric $acid_B$ receptors evolved earliest that have a common ancestor with the highest taxonomic rank (33213−Bilateria) compared to other subfamilies. CaSR group (CaSR, CaSR-likes, GPRC6A and taste receptors) was diverged from metabotropic glutamate receptors (GRM1-8) and RAIG receptors (GPRC5A, GPRC5B, GPRC5C, GPRC5D) with 1 and 0.98 TB values respectively. Within the CaSR group clade CaSRs and CaSR-likes were diverged from GPRC6A and taste receptors with 1 TB. Except TAS1R1 and TAS1R2, all CaSR group subfamilies have a common ancestor from taxonomy clade 7776-Gnathostomata. TAS1R1 and TAS1R2 were more specific than other CaSR group subfamilies that were evolved from 117571-Euteleostomi. Comparison analysis of branch lengths[20] among common species between CaSR, GPRC6A and taste receptors shows that CaSR subfamily is significantly more conserved than its closest subfamilies (Fig 2 B)

The higher diversity of CaSR-likes relative to CaSRs is reflected in the ML tree (Fig 2A). Branch lengths of CaSR-likes are longer in contrast to shorter branch lengths in CaSR. Longer branch lengths show that more variation, and thus divergence occurred in the CaSR-like clade.

Moreover, extensive gene duplication events occurred in this clade. For instance, rodents such as Dipodomys ordii (taxid:10020),Octodon degus**(**taxid:10160**)** and snakes such as Notechis scutatus(taxid: 8663) have more than a hundred receptors that match to CaSR profile. However, these matches include type 2 vomeronasal receptors (V2R) and type 2 vomeronasal receptor likes. Among mammals, V2R genes exhibit significant variation. While dogs, cows, and primates except prosimians do not have functional V2Rs**,** rodents, reptiles and fish have multiple intact V2Rs[21].Since these receptors do not have functional orthologs in mammals, separating them from functionally-equivalent CaSRs is crucial**.**

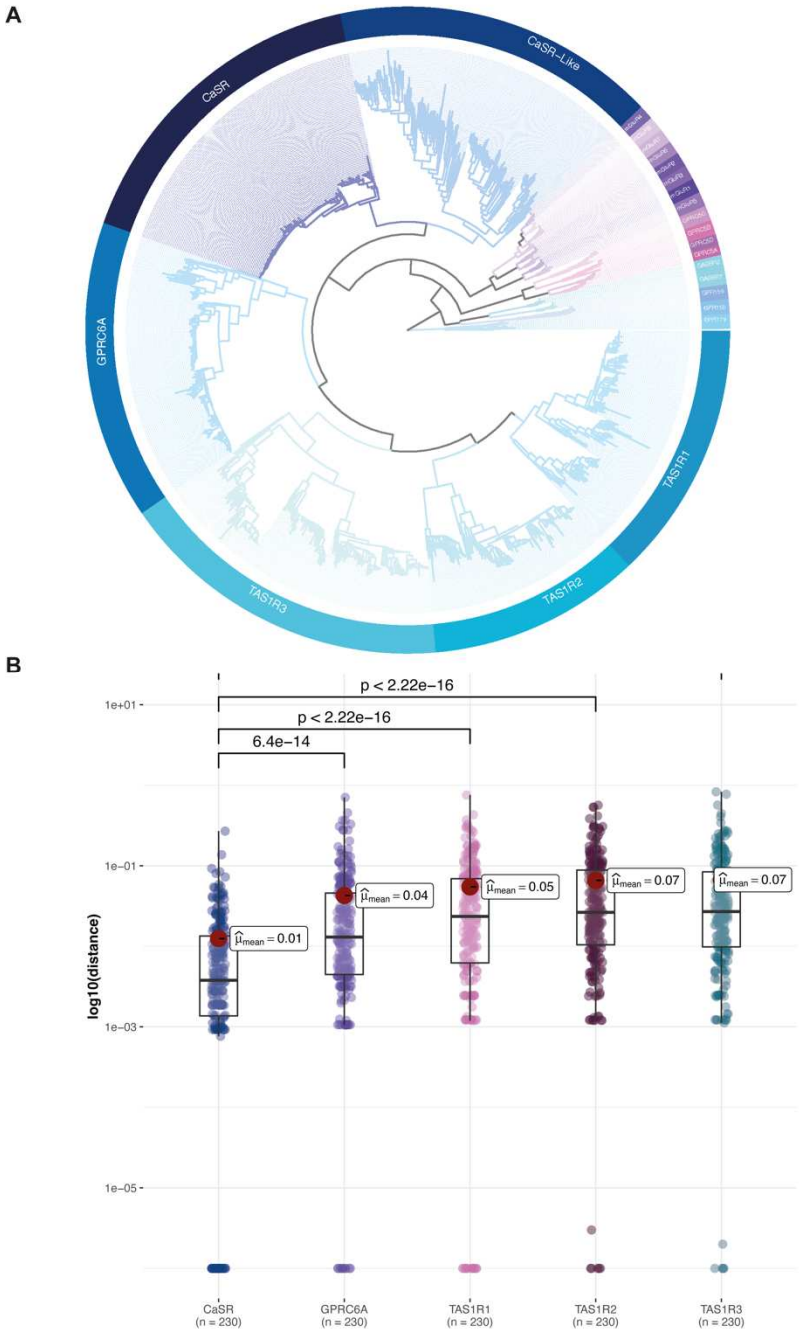

Figure 2: ***Evolution of Class C GPCRs****. **(A)** The maximum likelihood phylogenetic tree of Class C GPCRs, spanning representative species from each subfamily is shown. Subfamilies are represented as circular layers around the ML tree. All twenty-two Class C GPCR subfamilies are shown in the inner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

Subfamily-specific Profile HMMs to Obtain Orthologs

In the class C GPCR family, gene duplication events give rise to new specificity, and each duplicated gene with a new function is evolved by further speciation events and produce a set of orthologous sequences[15, 16]. Each subfamily of class C shares relatively conserved membrane-spanning region as well as a degree of variability underling functional differences. At the molecular level, residues that are responsible for certain functional characteristics such as ligand and coupling selectivity are called specificity-determining residues[15]. Conservation analysis from multiple sequence alignments can be used to find residues that are conserved in all subfamilies through evolution as well as specificity-determining residues that are only conserved in a subfamily and differ in other subfamilies. However, the success of this method depends on the sequences that are used to build alignments. Hence, it is vital to use functionally identical orthologs in the analysis.

The seven-transmembrane domains of class C GPCRs are used to build a class-specific general profile for this family (Pfam:7tm_3). However, this domain cannot be used to differentiate subfamilies further.

Moreover, excessive gene duplication events as seen in the CaSR-like clade requires precise phylogenetic analysis to differentiate CaSR and CaSR-like sequences. Also, subfamily specific profile HMMs are shown to be promising methods to detect protein sequences belong to a protein subfamily, as well as separation of homologs and non-homologs [22, 23].Therefore, we built subfamily-specific profile HMMs that match with all orthologs of a subfamily while excluding closely related like sequences (Fig 3).

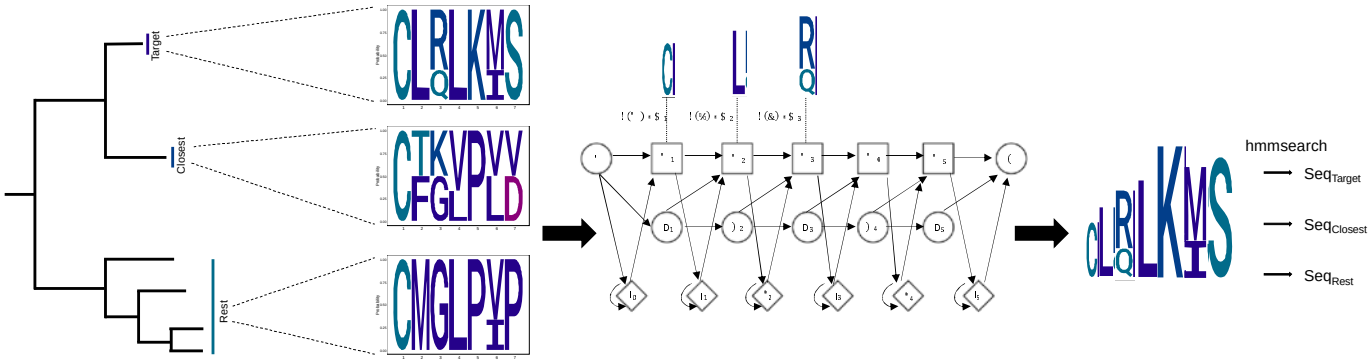

Figure 3: ***Subfamily Specific HMM Models.*** *Subfamily specific HMM model method. Based on the phylogenetic tree the target, the closest and the rest groups are determined. Representative amino acids in each group are selected, and their scores are calculated. Weights to scale the emission probability are calculated.*

We define the target family, its closest family (phylogenetic neighboring clade), and the rest based on the phylogenetic tree. We weighted 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. Minimum weight is given to the residues which are conserved both in the target subfamily and its closest clade. We tested our subfamily-specific profile HMMs' performance on an independent sequences retrieved UNIPROT dataset[24] and not seen during the training process. We assigned sequences to their corresponding subfamilies by following the same steps as NCBI dataset[25] used to build these models. We selected new taxons that were not in the NBCI dataset as test sequences. Our subfamily-specific profile HMMs correctly hit all members of a subfamily while they do not hit any protein from another subfamily (Table 1).

Table 1: Subfamily Specific Profile HMM's Performance

Subfamily HMM	Test Cases	Hits	Missed
CaSR	81	81	-
GPRC6A	62	62	-
TAS1R1	75	75	-
TAS1R2	21	21	-
TAS1R3	74	74	-

Specificity Determining Residues

CaSR is distinguished from other subfamilies of class C GPCRs by its oversensitivity to many substitutions that are caused either gain or loss of function mutations, because it maintains systemic calcium homeostasis and highly sensitive to a very slight change in extracellular Ca2+ concentrations[26]. Since CaSR is the most conserved and ancestral subfamily among the CaSR-likes, GPRC6A and TAS1Rs, it is reasonable to expect in some positions can be under the relaxation of existing purifying selection in any CaSR-likes, GPCR6A or TAS1Rs, but not in CaSR. On the other hand, at some positions the same amino acid remains functionally important in both subfamilies, and at others a position remains important in each subfamily but a different amino acid is favored in each duplicate.

To identify and order residues that differentiate a subfamily from its closest relatives, we employed multiple sequence alignment- and phylogenetic tree-based approaches. Specificity-determining residues that are conserved in a subfamily, but differ from its sister clade can be predicted by directly comparing ancestral family sequences and calculating their divergence scores (26). However, using multiple sequence alignments only does not discriminate between the number of substitution event. For example, a single substitution event in the common ancestor of bony fish clade of CaSR subfamily can be inherited to multiple descendants' sequences. Assessing this single event as independent events result in overcounting of these changes as if they are independent. Hence, the position is considered (i) to tolerate that particular amino acid and (ii) functionally less important. In contrast, a single evolutionary event might have been compensated by other substitutions in the same evolutionary node. Such a substitution might not be tolerated in the other clades of the subfamily.

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

For CaSR group(CaSR, GPRC6A and TAS1Rs), we identified and order residues by specificity which differentiate a subfamily from others by using an adaption of functionally divergent residues method[27] along with an adaption of PHACT method[17]. We calculated probability of

each amino acid at each node of the CaSR-group phylogenetic tree by ancestral sequence reconstruction (Fig 4 A). Starting from the root of the tree, we identified each substitution event and at which subfamily node that event happened. Counting the number of independent substitution events in a subfamily clade 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 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 specificity of a residue because it diverges the target group from its closest, sister clade. The details and the algorithm are given in materials and methods (Algorithm 3).

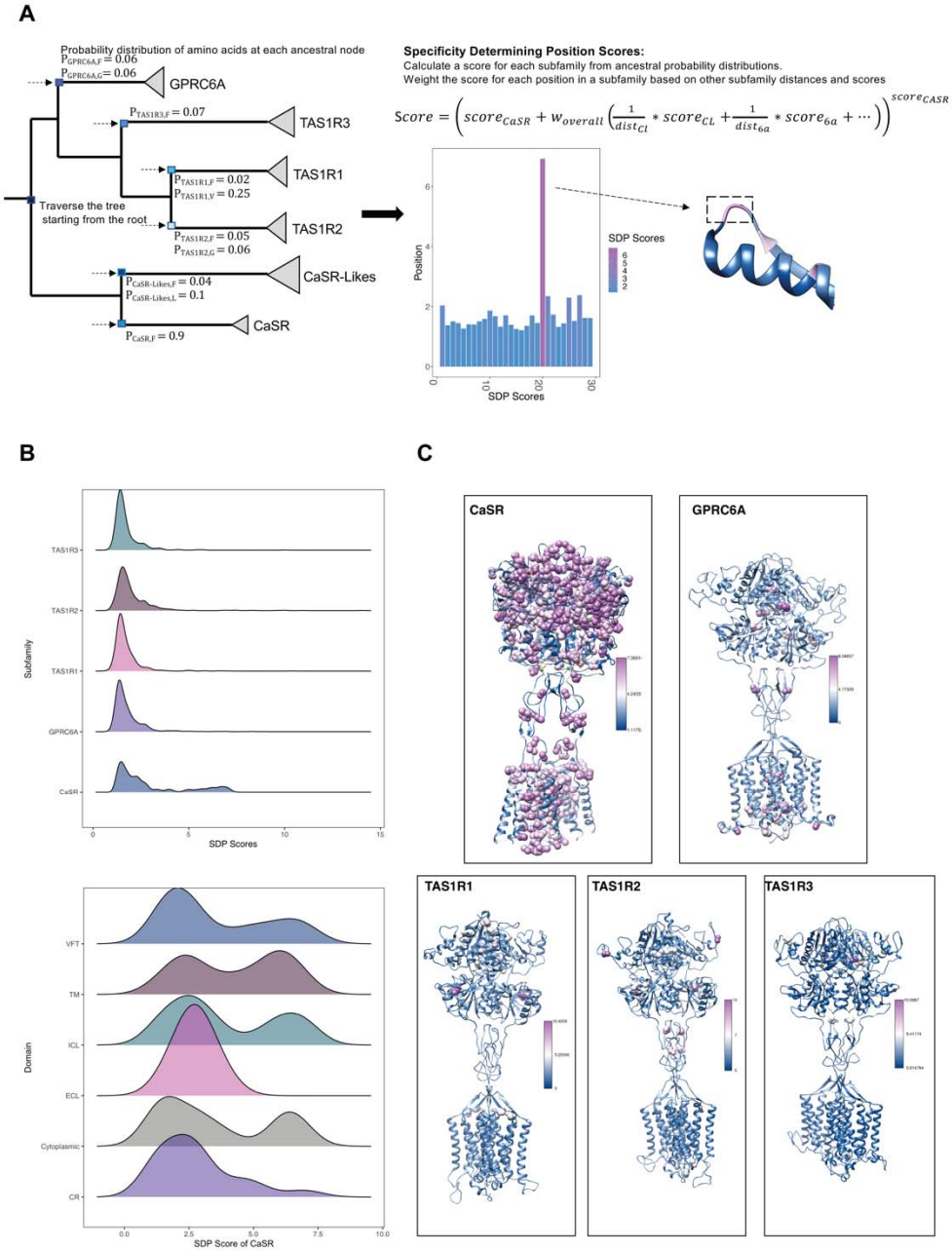

Figure 4: ***Specificity Determining Position Scores. (A)*** *Calculation of SDP scores uses the phylogenetic tree, probability distribution of amino acids at each ancestral node.* ***(B)*** *SDP score distributions of each subfamily are shown.* ***(C)****Cryo-EM structure of human CaSR bound with* Ca^{2+} *and L-Trp (PDB:7DTV) and homology models of*

GPRC6A, TAS1R1, TAS1R2 and TAS1R3 are colored based on SDP scores. Residues with high SDP score (above 5.0) are shown as spheres.

We calculated specificity scores for each CaSR, GPRC6A and TAS1Rs. Specificity score distributions show that CaSR subfamily have more specific residues compared to other subfamilies (Fig 4 B). On the VFT domain, specific residues are clustered different regions. We found a cluster of specific residues on the interdomain cleft between LB1-LB2 that is the L-amino acid binding site in other class C GPCRs[3]. It suggests that this region is the primary Ca2+ binding site in CaSR consistent with[14]. We found two different clusters of specific residues on the ECD. First cluster was on the LB1 domain and on the LB1-LB1 dimer interface. LB1 domain plays a role in anchoring ligands and initiating domain twisting by conformational changes at the interface between LB1 regions[3, 5]. The second cluster was found at the cytosolic side of the LB2 and at the interface between LB2-CRD where Ca2+ ions are bind[3-5]. Interaction between LB2 subunits are required for CaSR activation that propagates to large-scale transitions of the 7TMDs[3, 5]. Specific residues on the LB1 domain, LB1-LB1 dimer interface and LB2-CRD interface indicate that they provide the structural conformational changes upon ligand binding to the interdomain cleft. Mutations located on these regions are associated with loss and gain of function mutations (Fig 6)[2]. Other specific residues are found on the CR, ECL2 and TM domains. On the ECL2 acidic residues D758 and E759 are specific to CaSR. The intersubunit electrostatic repulsion between the ECL2 regions could facilitate the activation of CaSR[3, 5]. In the agonist+PAM bound state the ECL2 is moved by the interaction among E759, W590, and K601. Deletion of D758 and E759, and single mutations of K601E and W590E disrupts the CaSR activity, however Δ758–759 mutant was expressed at the cell surface with the comparable levels to that of WT, while W590E and K601E mutants were expressed on the cell surface lower than the WT level[3]. We found that residues W590 and K601 are not specific to CaSR. The TM domains of two protomers of CaSR come into close proximity upon receptor activation[5].

The orientation of the TM5-TM6 dimer in the CaSR distinguishes it from other Class C receptors such as mGluR and $GABA_B$ receptors, which results in its inactive conformation[9]. The interaction between TM4-5 of each subunit in the inactive state is essential[14], while the interaction between TM6-TM6 is crucial for the active state[3, 8, 14]. The structural findings and the presence of CaSR specific residues on each TM domain suggest that CaSR is specialized in both dimerization and ligand binding. Specific residues on TM domain guarantees the correct orientation for activation upon ligand binding and inactive conformation otherwise. Interactions between the domains and ligand-receptor are quite sensitive that slight changes cause

malfunctions in the receptor. On the other hand, GPRC6A and taste receptors are more prone to acceptable substitutions and they are not very specialized to respond a single ion. GPRC6A and taste receptors are activated by a broad spectrum of ligands[28, 29]. Even though the ligand of GPRC6A is controversial in the literature, multiple ligands such as osteocalcin (Ocn), testosterone, basic amino acids and cations such as L-Arg, L-Lys, L-Orn, calcium, magnesium, and zinc are suggested to bind GPRC6A[29]. Taste receptors bind to different ligands including sugar, L- and D-amino acids, sweet proteins, and artificial sweeteners[30].

On the TM region we also find CaSR specific cholesterol recognition/interaction amino acid consensus (CRAC) motif (L783,F789,S820) that is defined by the consensus (L/V)X1–5YX1–5(R/K) and is often present at junctions between membrane- and cytosol-exposed domains and shown in GRM2 receptor[31]. Phylogenetic analysis shows that TAS1R3 evolved earliest (7776 Gnathostomata) among TAS1Rs, TAS1R1and TAS1R2 subfamilies have common ancestor 117571 Euteleostomi. TAS1R3 forms heterodimers with TAS1R1 and TAS1R2[28, 30, 32]. Interactions between the cytosolic terminus of the extracellular CRD is needed for T1R3 dimerization. TAS1R1 and TAS1R2 recognize a broad spectrum of L-amino acids that bind to the intercleft between LB1-LB2 and induce the positional shift of the CRD regions, however T1R3 loses the corresponding function[32]. Our analysis showed that TAS1R1 have specific residues on LB1, LB2 and extracellular loop regions. Also, TAS1R2 has specific residues on LB1,LB2 and CR domains. On the other hand, in TAS1R3, we found specific residues only on the LB1 and one on the CR domain. Since LB1-LB2 domains create a cavity for ligand binding, specific residues on LB1-LB2 domains of TAS1R1 and TAS1R2 may contribute to domain transformation upon ligand binding. However, the number and distribution of specific amino acids suggest that taste receptors are not under selective pressure as CaSR.

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 gain or loss of function mutations (Fig 3C). However, predicting the functional consequence of a substitution is challenging. Evolutionary conservation of a residue among subfamilies might reflect the common structural constraints, but it does not distinguish between loss and gain of function mutations (i.e., LoF and GoF, respectively). In addition, at some positions substitution to different amino acids causes either loss or gain of function mutations[21]. We hypothesized that “activating” mutations are more likely to be tolerated in the neighboring clades such as GPRC6A and TAS1Rs and not in CaSR whereas, in general, loss-of-function (inactivating) mutations are not tolerated in the larger clade of these receptor subfamilies. To test this hypothesis whether we can discriminate between GoF

and LoF mutations in CaSR, we applied a tree boosting machine learning algorithm, XGBoost[33] that linked multiple features such as conservation scores, physico-chemical properties of amino acids and domain information.

We used sequence-based features, identity scores from multiple sequence alignments, physico-chemical properties of amino acids, and domain information as input features to train our model (Fig 5 A). Since we calculated our feature values from the multiple sequence alignments, we divided our dataset into training, validation and test datasets before we created feature matrices to prevent information leakage. We performed 50 replications with different random splitting of datasets to obtain a more robust model performance.

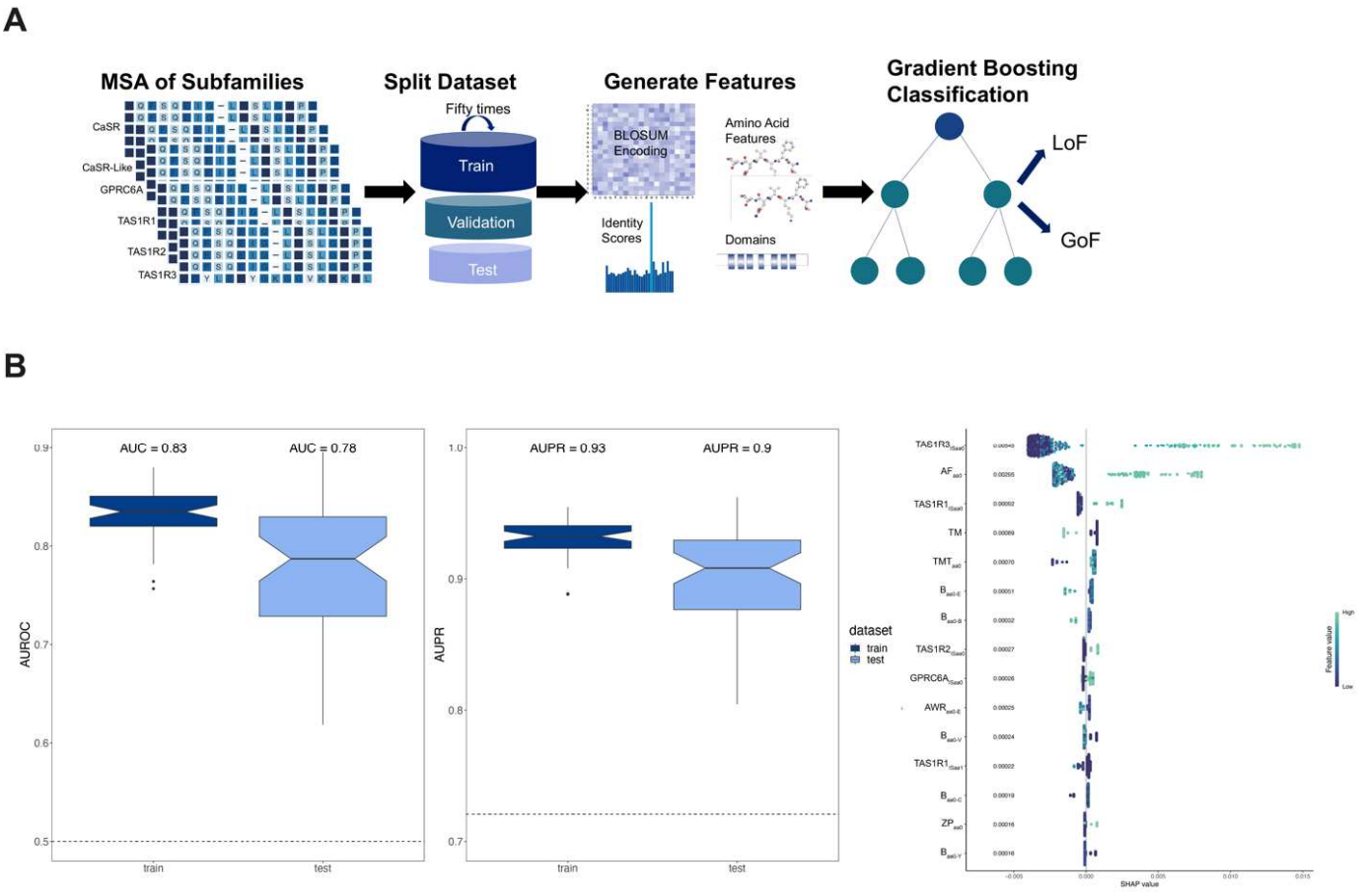

Figure 5: ***Gradient Boosting Trees Machine Learning Approach to Predict the Mutation Types in CaSR.*** ***(A)****Model architecture. We used MSA of CaSR,CaSR-likes, GPRC6A and TAS1Rs to generate features as well as amino acid physico-chemical features and domain information. We performed 50 replications.* ***(B)*** *The performance and feature importance of XGBoost algorithm. AUROC and AUPR values of 50 replications are shown. Average AUC levels of 50 replications are 0.83 and 0.78 for the train and test respectively. 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*

342 *acid, AF: average flexibility, TMT: TM tendency, ZP: Zimmerman polarity, B:BLOSUM62,AWR:atomic weight*
343 *ratio,TM:transmembrane domain*

344 *Table 2: Model's predictions for the new CASR gain and loss of function mutations from the literature. The correct*
345 *predictions are indicated by a star symbol (*) next to them.*

Mutation	Cause	Prediction
p.I857S[34]	hypocalcemia	gain-of-function*
p.Y825F [35]	hypocalcemia	gain-of-function*
p.P393R [36]	hypercalcemia	loss-of-function*
p.C60G[37]	hypercalcemia	loss-of-function*
p.D99N[38]	hypercalcemia	loss-of-function*
p.T186N[39]	hypocalcemia	loss-of-function
p.A840V[24]	hypocalcemia	gain-of-function*
p.S448P[40]	hypercalcemia	loss-of-function*
p.L696V[41]	hypocalcemia	gain-of-function*
p.D433Y[42]	hypercalcemia	loss-of-function*
p.S147L[43]	hypercalcemia	loss-of-function*
p.D398N[44]	hypercalcemia	loss-of-function*
p.K805R[45]	hypercalcemia	gain-of-function
p.C60Y[46]	hypercalcemia	loss-of-function*
p.S820N[47]	hypocalcemia	loss-of-function
p.L606P[48]	hypercalcemia	loss-of-function*
p.H41R[49]	hypercalcemia	gain-of-function
p.A110D[50]	hypercalcemia	gain-of-function
p.I139T[51]	hypocalcemia	gain-of-function*
p.Q164R[52]	hypercalcemia	loss-of-function*
p.T699N[53]	hypercalcemia	gain-of-function
p.R701G[53]	hypercalcemia	loss-of-function*
p.T808P[53]	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. ROC curve is a graphical plot that shows the false positive rate versus the true positive rate for different threshold values between 0.0 and 1. A PR curve is a plot of the precision and the recall for different threshold values and it is useful for imbalance datasets. We used the areas under the ROC and PR curves (i.e., AUC and AUPR, respectively) to compare the performances of the model on the train and test datasets for 50 replications. Higher AUC and AUPR values are associated with better performance. AUC and AUPR over all replications were shown in (Fig 5 B). Our average AUC values for training and test among 50 replications are 0.83 and 0.78 (Fig 5 B). Our average main AUPR values for training and test among 50 replications are 0.93 and 0.9 respectively (Fig 5 B). After we reported our algorithm performance, we trained our algorithm with the whole dataset. We tested our algorithm with new test cases from literature (Table 2). Additionally, we categorized amino acids that are observed in the CaSR MSA as neutrals. To date, no pathogenic substitution has been reported in the literature for these amino acids that we identified as neutral. We visualized all predictions in the form of a heatmap for every other amino acids at each position until the disordered region (position 892) of the human CaSR (Fig 6 A). We mapped known CaSR loss and gain of function mutations on the cryo-EM structure of human CaSR bound with Ca^{2+} and L-Trp (PDB:7DTV (3)) (Fig 6 B). There is a tendency that loss-of-function mutations are on the outer-core regions, while gain-of-function mutations are on the inner-core regions. In the heatmap we observed a similar prediction pattern that gain-of-function predictions are mostly in the inner-core regions. SHAP (SHapley Additive exPlanations) values provide a way to decode the inner workings of a machine learning model like XGBoost. These values calculate the average contribution of each feature to the overall prediction, taking into account any interactions between the features. Based on the SHAP values, the conservation scores of human CaSR amino acids in other subfamilies play a significant role in the model's prediction, as shown in Figure 5B. If the amino acid is also conserved in GPRC6A and taste receptors (in fact conservation score in TAS1R3 has the highest contribution), the model predicts a substitution of that amino acid as loss-of-function. Another important feature is the domain of the amino acid. Our findings indicate that if the amino acid is located in the TM domain, a substitution would result in a gain-of-function mutation.It is known that the majority of gain of function mutations are located in the TM domain, as shown in Figure 6B. The presence of certain amino acids on the TM domain of CaSR suggests that they play a crucial role in its activation mechanism. Even though substituting those amino acids might be acceptable in

GPRC6A and taste receptors, they might lead to the lock of TM domains and result in the overactivation of CaSR.

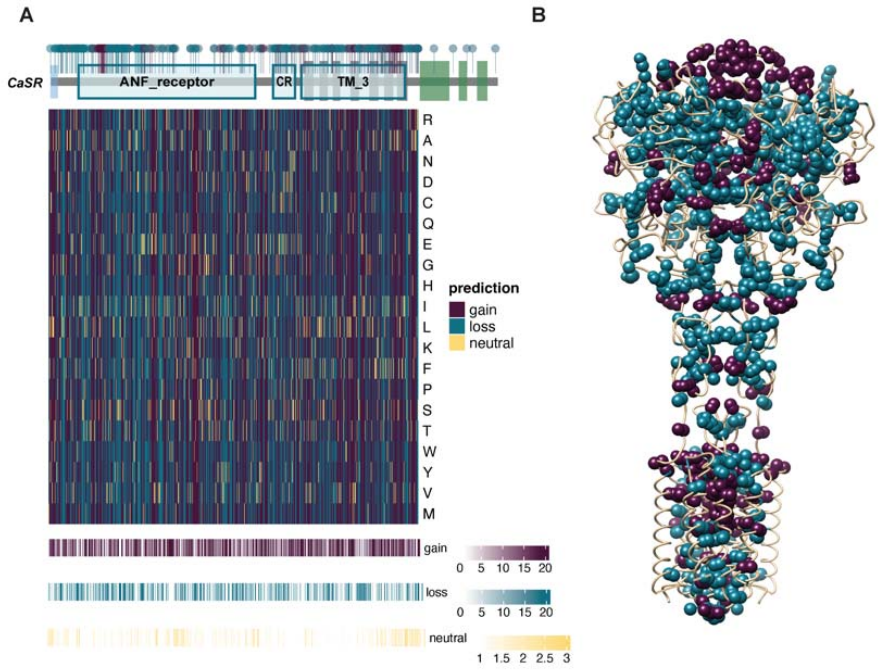

Figure 6: ***All Possible Amino Acid Substitution Predictions. (A)Visualizing*** *the precision of our XGBoost model. The heatmap displays the XGBoost model's predictions for each of the 20 amino acids at every position except disordered regions (892-1078) within the human CaSR.* ***(B)*** *Mutations on human CaSR structure. Loss and gain of function associates mutations are shown on the cryo-EM structure of human CaSR bound with Ca2+ and L-Trp (PDB:7DTV) as blue and red spheres respectively.*

Discussion

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

High evolutionarily conservation and specificity of CaSR in contrast to closest subfamilies are reflected in specificity determining position (SDP) score analysis. CaSR has specific residues clustered on different regions of the receptor. They are located on Ca2+ and L-Trp binding sites

on the VFT, as well as on the dimerization sites between two sub-units of the homodimer. Specific residues on the dimer interfaces indicate that dimerization maintained by interactions between different subunits is required for ligand binding and correct activation of the CaSR. Ca2+ ions binding and interactions between LB2-CR domains and conformational changes in LB1 domain were suggested that they are required to activate CaSR[3-5]. Mutational analysis at some positions on LB1 domain have been shown to reduce the effect of Ca2+-stimulated intracellular Ca^{2+} mobilization in cells[3, 5]. In contrast, substitutions caused negative charge neutralization on the ECL2 result in prompting the activation of CaSR[5]. Our results suggest that residues with low SDP scores on any domain are required for common activation mechanism since they are conserved across functionally different receptor subfamilies. However, residues with high SDP scores cause malfunctions in the CaSR. Any substitution in a residue with high SDP score might either cause over or less activation. Deep mutational scanning approaches or new methods that simultaneously profile variant libraries[54] are needed to provide further evidence to functionally assay all possible missense mutants.

To predict the functional consequence of a mutation in human CaSR, we used Extreme Gradient Boosting (XGBoost) method. XGBoost is able to perform well on small datasets by incorporating variety of regularization methods to control the model complexity, which helps to prevent overfitting. We have a small and unbalanced dataset in that the number of gain-of-function mutations was very low, therefore it is prone to overfit. To prevent overfitting while achieving high predictive performance, we used a simple method along with regularization parameters. Moreover, we tried to keep the ratio between the number of loss-of-function and the number of gain-of-function mutations for training and test sets as close as possible. To get a robust performance, we repeated the train-validation-test splitting procedure fifty times. To increase the predictive performance, we could use a more complex methods such as deep learning, however they require larger datasets. Studies that used deep learning or ensemble methods for similar assessment are different in terms of prediction in which they predict the type of a mutation as only pathogenic or neutral [55-58]. Even though there are number of mutations of human CaSR in the Clinvar, the functional consequences of most of them are not known.

Given the constraints of the small dataset and limited additional data, we carefully selected and processed the features for our model's training. Features that are used to train a machine learning model heavily determine the performance of the model. The more features we use, the more information the model has to learn from, which can lead to improved predictive performance. However, having too many features can also lead to overfitting. Moreover, the quality of the features is more important than the quantity. One important evolutionary process

that can affect the functional consequence of a substitution is co-evolution. From the multiple sequence alignment of CaSR proteins, we manually selected six positions,p.180,p.212,p.228,p.241,p.557 and p.883, that are in our dataset and co-evolved. We masked the co-evolved amino acids from the MSA and performed train-validation-test splitting procedure fifty times again. Our average AUC values for training and test among 50 replications were 0.83 and 0.77 respectively, and average AUPR values were 0.93 and 0.89. Despite not experiencing an improvement in performance, we found that the amino acid changes p.I212T, p.F180C, and p.I212S were now predicted to cause loss of function, contrary to their previous prediction of causing gain of function. We cannot accurately assess the impact of co-evolution on performance because there is a lack of effective tools for identifying co-evolved positions and our dataset contains only a limited number of co-evolved positions, but we anticipate that it is an important feature to differentiate gain and loss of function.

We built subfamily-specific profile HMMs to get all functionally-equivalent orthologs while excluding other proteins. To generate these HMM models, we manually decided target, closest and rest groups based on the phylogenetic tree of CaSR group. Based on the nature of a phylogenetic tree, selection of these groups is changed, so that this process can be further automated. We did not anticipate our specific models to match any receptor from other classes of GPCRs, since they are evolutionarily more distant to CaSR group. We expect that our subfamily specific profile HMMs can be used to obtain orthologs in different protein families for the upcoming genomes. They can be particularly useful for studying protein families with many duplications and orphan protein families, where it can be difficult to identify true members. These models are particularly important to avoid computationally expensive and expertise-required phylogenetic tree reconstruction and analysis.

Materials and methods

Class C Proteins and Their Domain Architectures

478 complete eukaryotic proteomes were downloaded from NCBI genomes website(https : //ftp.ncbi.nlm.nih.gov/genomes/archive/old ref seq/) in 2018. hmmsearch of HMMER software[59](http://hmmer.org/) was run for each proteome against Pfam 7TM3 profile[60]. Sequences with significant 7TM3 hit based on hmmsearch results (above the default threshold) were compiled from proteomes. hmmscan of HMMER software[59](http://hmmer.org/) was run for these sequences against Pfam-A 32.0 database[60]. Based on the results of hmmscan, the longest isoform was taken and saved in a separate file named by taxonomic id, however canonical sequences were obtained for human (based on given canonical proteins in UniProt

website[61]). Because plants do not have GPCRs, plants were eliminated from the analysis. For single isoform sequences of each proteome a BLAST database was built[62].

Subfamily Definition and Subfamily Specific Models

Each protein sequence of each taxon was queried through BLASTP against each prepared BLAST database[62]. reciprocal mutual best hits of each human class C GPCR were collected in a file named gene id. reciprocal mutual best hits of each class C GPCR and remaining human class C GPCRs were collected and 7TM domains of these sequences were taken based on hmmscan results (Longest sequence which hit the 7TM3). Sequences were aligned using MAFFT v7.221 E-INS-I algorithm with default parameters[63]. Maximum likelihood based phylogenetic tree (ML tree) of each subfamily of class C GPCR was built using RAxML version 8.2.12 with automatic protein substitution model selection (PROTGAMMAAUTO) and 100 rapid bootstrapping parameters[64]. Most common lowest taxonomic level was added to the phylogenetic tree with ETE toolkit[65]. Based on the phylogenetic tree, sequences belong to the corresponding subfamily were taken and an profile HMM was built. Subfamily Assignment The process begins by scanning each sequence with a 7TM3 domain against profile Hidden Markov Models (profile HMMs). After the sequence is scanned, the subfamily is determined based on three conditions: (1)The maximum score value of the hmmscan must belong to the given subfamily. (2) E-value is a measure of the significance of a match in a database search and the lower the E-value, the more significant the match is. The E-value of the sequence must be the lowest. (3)The sequence must belong to the most common highest taxonomic level of the given subfamily. Taxonomic level refers to the classification of an organism within a biological classification system. If a sequence meets these three conditions, it is assigned to the corresponding subfamily. After this, the full length sequences of each subfamily were then aligned using the MAFFT v7.221 algorithm[63] and trimmed using the gappy-out method of the trimAl tool[66].

Paralog Filter

There were a number of duplications in CaSR subfamily. For example, Dipodomys ordii has 116 CaSR sequences. To reduce the number of seqeunces,human CaSR and other human class C GPCR proteins sequences compiled with CaSR seqeunces of each taxon, and and aligned with MAFFT v7.221 auto algortihm[63], and the gappy-out method of the trimAl tool was used to trim the multiple sequence alignments (MSA)[66]. ML tree was built using RAxML-NG v0.9.0 with ML tree search and bootstrapping (Felsenstein Bootstrap and Transfer Bootstrap) parameters[67]. Based on the ML tree, proteins that were diverged from the common ancestor of the human

CaSR clade were classified as CaSR-likes. Proteins that were clustered with the human CaSR were accepted as CaSRs. After we assigned all proteins to their subfamilies, we built final ML trees for CaSR,GPRC6A, and TAS1Rs. We added human CaSR sequence was added to GPRC6A and TAS1Rs subfamilies, and human GPRC6A sequence was added to CaSR subfamily as an outgroup. We aligned each subfamily sequences with MAFFT v7.221 einsi algortihm[63] and built the ML trees by using RAxML-NG v0.9.0 with FTT model parameter[67]. We labeled the duplications at each node on the ML trees. Based on the duplications, we manually checked the trees and removed a clade that was a subset of its sister clade by using ETE toolkit[65]. We took each branch and node length from leaf to root of the tree by using common species in all CaSR, GPRC6A and taste receptor trees to calculate subfamily conservation by using Welch's t-Test by using ggstatsplot package[20].

Subfamily Specific Profile HMMs

[63].After we took all receptors from CaSR, CaSR-like, GPRC6A, and taste receptors, we aligned them by using MAFFT v7.221 auto algortihm[63]. For each subfamily we removed the positions from the multiple sequence alignment (MSA) that correspond to a gap in the human receptor. Then, we divided the MSA into subfamily alignments. We generated a HMM from the gap removed alignment of each subfamily, and we added weight to the emission probabilities of the HMMs. To calculate emission probability weights, based on the maximum likelihood phylogenetic tree (ML tree) we defined the target, the closest and the rest groups. We took the closest node as the closest group and other nodes as the rest. According to that we have five different scenarios:

- CaSR is the target group, CaSR-likes are the closest group, and GPRC6A and taste receptors, (TAS1Rs) are the rest.
- GPRC6A is the target group, TAS1Rs are the closest group, CaSR and CaSR-likes are the rest.
- TAS1R1 is the target group, TAS1R2 is the closest group and TAS1R3 is the rest.
- TAS1R2 is the target group, TAS1R1 is the closest group and TAS1R3 is the rest.
- TAS1R3 is the target group, TAS1R1 and TAS1R2 are the closest group and GPRC6A is the rest.

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; 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

Input: *Types for each position k (k=1,…,K),* $type_k$*; initial score for each position k of type t,* $score_k^t$*; 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 II positions,* c_2*.*

Weight of Type I positions

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

  $weight_{p_1} = \frac{score_{p_1}^1}{\min_{l=1,\dots,n_1}(score_l^1)}$

Weight of Type IV positions

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

  $weight_{p_2} = \frac{score_{p_2}^2}{\max_{l=1,\dots,n_2}(score_l^2)} mean\left(weight_{p_1}\right)_{p_1\in\{1,\dots,n_1\}}$

Weight of Type III positions

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

  if target is CaSR

7 $$c_1 = \frac{\min\left(weight_{p_2}\right)_{p_2 \in \{1,\dots,n_2\}}}{2}$$

8 *else*

9 $$c_1 = \min\left(weight_{p_2}\right)_{p_2 \in \{1,\dots,n_2\}}$$

4 $$weight_{p_3} = \frac{score_{p_3}^3}{\max_{l=1,\dots,n_3}\left(score_l^3\right)} c_1$$

Weight of Type II positions

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

8 $$weight_{p_4} = \frac{score_{p_4}^4}{\max_{l=1,\dots,n_4}\left(score_l^4\right)} c_2$$

528

529 **Subfamily Specific Position Scores**

530 From the alignment we used to make subfamily specific profile HMMs, we randomly selected
531 264 CaSR like sequences (same number of sequences as CaSRs) and took all CaSR (264
532 proteins), GPRC6A (242 proteins) and TAS1Rs (TAS1R1 has 210, TAS1R2 has 173 and
533 TAS1R3 has 273 proteins). We built an ML tree by using IQ-TREE multicore version 2.0.6[68]
534 with automatic model selection[69] (-m MFP) and ultrafast bootstrap[70] (-bb 1000) parameters.
535 For CaSR, GPRC6A, and TAS1Rs, we removed the positions from the multiple sequence
536 alignment that correspond to a gap in the human receptor respectively. By using gap removed
537 alignments and the ML tree, we did ancestral sequence reconstructions for each subfamily with
538 IQ-TREE multicore version 2.0.6 with -m JTT+R10 model parameter[68]. We showed specific
539 residues that have a SDP score higher than 5, on the structures. We used cyro-EM structure of
540 CaSR (PDB:7DTV) and Swiss models[71] for GPRC6A and taste receptors since they do not
541 have experimental structures. To visualize structures and residues we used UCSF Chimera
542 tool[72].

543 We calculated SDP scores by a method extended from[27] by considering the phylogenetic
544 trees and a phylogeny-based scoring approach, adjPHACT, based on the methodology of
545 PHACT algorithm. The details of how we compute SDP score for any position k can be found in
546 Algorithm 3. PHACT computes the tolerance for each amino acid for the query specie which is
547 human by using a tree traversal approach. By checking the probability differences, PHACT
548 detects the location of amino acid substitutions and compute weighted summation of positive
549 probability differences based on the distance between the node of change and human. On the
550 other hand, here we aim to determine the acceptability of each amino acid per subfamily. To

achieve this, we modify PHACT by starting the tree traversal from the root node and eliminating the node weighting approach. At the end, we have a probability distribution per position for each subfamily which is computed by considering the independent events. Again, we determine the representative amino acid for target subfamily by picking the most frequently observed amino acid and its adjPHACT score. For the remaining subfamilies, we keep the adjPHACT score of the representative amino acid of the target. Then, similar to [27] we check whether the same amino acid is conserved across all subfamilies. On the other hand, our approach differentiates from[27] in terms of considering multiple subfamilies and using adjPHACT scores which employ phylogenetic trees and ancestral reconstruction probabilities. In our approach, we compute the contribution of each subfamily to the SDP score by checking whether the representative amino acid of target has a high adjPHACT score in that subfamily (line 1). In the final SDP score for any position k is computed by considering the distance between target and other subfamilies (which is computed by considering the distance between root nodes), the conservation level of the target subfamily in terms of independent amino acid alterations and the individual score coming from each subfamily (line 3).

ALGORITHM 3: SDP SCORE FOR POSITION "K"

Input: *Amino acid with the highest adjPHACT score in the target group, aa; the adjPHACT score of aa for target,* P_{aa}^{T}*; adjPHACT score of aa for other subfamilies i=1,…,n,* P_{aa}^{i}*; distance between target subfamily and subfamily i,* d_i*.*

1 *Compute score for each subfamily i,*

$$S_i = \exp(1) - \exp\left(P_{aa}^{i}\right).$$

2 *The overall weight,*

$$\omega = 1 - \max\left(P_{aa}^{i}\right).$$

3 *The SDP score for position k,*

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

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 MAFFT v7.221 einsi algorithm[63]. We built the ML tree by RAxML-NG-0.9.0 with the model JTT and transfer bootstrap expectation –bs-

metric fbp, tbe parameters[67]. We merged the ML trees of CaSR, GPRC6A and taste receptors by checking clades by using ETE toolkit [65].

Machine Learning

Dataset and Feature Preparation

To predict the consequence of a substitution in human CaSR, we used a gradient boosting-based machine learning algorithm, XGBoost[33]. We used XGBoost library for R[73] to train our model. We selected total of 337 loss and gain-of-function mutations from the literature[2] to train our model. Since we used conservation scores as features to train our model, we divided subfamily alignments and mutations randomly as 80% training and the remaining 20% test data before creating feature matrices to prevent information leakage. 25% of the training data was randomly picked as the validation data five times for cross validation. For each dataset split we used the sklearn train test split model with stratify option to keep loss-of-function to gain-of-function ratio almost the same in the datasets[74]. We calculated the conservation score of the reference amino acid and the substituted amino acid in human CaSR in each subfamily. The reference and the substituted amino acids were represented BLOSUM62 encoded matrices. Amino acid physico-chemical feature values Zimmerman polarity[75], average flexibility[76], Dayhoff[77], average buried area[78], Doolittle hydropathicity[79], atomic weight ratio[80], molecular weight, and bulkiness[75] from ProtScale database[81]; and domain information of the reference amino acid were used as other features. We normalized the physico-chemical feature values prior to model training. We repeated the whole random dataset splitting and feature preparation procedure 50 times to obtain more robust results.

Model Selection and Parameter Tuning

We picked the model parameters for each replication by applying a 5-fold cross-validation technique on the training set. We tuned the model parameters step-by-step using the same validation sets for each parameter to decrease the time complexity. We used the following order of model parameters, so that the parameter that has the highest impact on model outcome was tuned first: Eta and nrounds, gamma, maxdepth, subsample, colsample bytree, min child weight, lambda, alpha. We selected the maxdepth as 2, the minimum maxdepth value to prevent overfitting. We chose eta, gamma, colsample bytree, subsample, min child weight from 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 respectively. We selected regularization parameters lambda and alpha from the set 0, 1e-4, 1e-3, 1e-2, 1e-1, 1, 10, 100. We set nrounds parameter as 200.

Performance Metrics

We used the area under the receiver operating characteristic curve (AUROC) and the area under the precision-recall curve (AUPR) to evaluate the performance of our prediction model. AUROC and AUPR are performance measures that are widely used to evaluate the performance of binary classification problems. The higher the AUROC and AUPR, the better the model distinguishes classes. To understand how our model makes predictions, we used SHAP (SHapley Additive exPlanations) values. SHAP values provide an estimate of the contribution of each feature to the prediction made by the model[82]. We calculated SHAP values for our final model trained by all samples by using R shapviz package[83].

Predictive Performance

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

Acknowledgments

This study was supported by EMBO Installation Grant no:4163 funded by TÜBİTAK (to OA). Ogun Adebali is supported by the BAGEP program of the Science Academy - Türkiye, and the TÜBA-GEBİP program of the Turkish Academy of Sciences.

References

1. Cook, A.E., et al., *Biased allosteric modulation at the CaS receptor engendered by structurally diverse calcimimetics.* British journal of pharmacology, 2015. **172**(1): p. 185-200.
2. Gorvin, C.M., *Molecular and clinical insights from studies of calcium-sensing receptor mutations.* J Mol Endocrinol, 2019. **63**(2): p. R1-R16.
3. Chen, X., et al., *Structural insights into the activation of human calcium-sensing receptor.* Elife, 2021. **10**.
4. Geng, Y., et al., *Structural mechanism of ligand activation in human calcium-sensing receptor.* Elife, 2016. **5**.

5. Ling, S., et al., *Structural mechanism of cooperative activation of the human calcium-sensing receptor by Ca(2+) ions and L-tryptophan.* Cell Res, 2021. **31**(4): p. 383-394.
6. Wootten, D., et al., *Mechanisms of signalling and biased agonism in G protein-coupled receptors.* Nat Rev Mol Cell Biol, 2018. **19**(10): p. 638-653.
7. Zhang, C., et al., *Structural basis for regulation of human calcium-sensing receptor by magnesium ions and an unexpected tryptophan derivative co-agonist.* Sci Adv, 2016. **2**(5): p. e1600241.
8. Gao, Y., et al., *Asymmetric activation of the calcium-sensing receptor homodimer.* Nature, 2021. **595**(7867): p. 455-459.
9. Park, J., et al., *Symmetric activation and modulation of the human calcium-sensing receptor.* Proc Natl Acad Sci U S A, 2021. **118**(51).
10. Wen, T., et al., *Structural basis for activation and allosteric modulation of full-length calcium-sensing receptor.* Science Advances, 2021. **7**(23): p. eabg1483.
11. Mun, H.-C., et al., *A double mutation in the extracellular Ca2+-sensing receptor's venus flytrap domain that selectively disables L-amino acid sensing.* Journal of Biological Chemistry, 2005. **280**(32): p. 29067-29072.
12. Zhang, C., et al., *Identification of an L-phenylalanine binding site enhancing the cooperative responses of the calcium-sensing receptor to calcium.* Journal of Biological Chemistry, 2014. **289**(8): p. 5296-5309.
13. Zhang, Z., et al., *Three adjacent serines in the extracellular domains of the CaR are required for L-amino acid-mediated potentiation of receptor function.* Journal of Biological Chemistry, 2002. **277**(37): p. 33727-33735.
14. Liu, H., et al., *Illuminating the allosteric modulation of the calcium-sensing receptor.* Proceedings of the National Academy of Sciences, 2020. **117**(35): p. 21711-21722.
15. Chagoyen, M., J.A. Garcia-Martin, and F. Pazos, *Practical analysis of specificity-determining residues in protein families.* Brief Bioinform, 2016. **17**(2): p. 255-61.
16. Studer, R.A., B.H. Dessailly, and C.A. Orengo, *Residue mutations and their impact on protein structure and function: detecting beneficial and pathogenic changes.* Biochemical journal, 2013. **449**(3): p. 581-594.
17. Kuru, N., et al., *PHACT: Phylogeny-Aware Computing of Tolerance for Missense Mutations.* Mol Biol Evol, 2022. **39**(6).
18. Pin, J.P., T. Galvez, and L. Prezeau, *Evolution, structure, and activation mechanism of family 3/C G-protein-coupled receptors.* Pharmacol Ther, 2003. **98**(3): p. 325-54.

19. Harpsoe, K., et al., *Structural insight to mutation effects uncover a common allosteric site in class C GPCRs.* Bioinformatics, 2017. **33**(8): p. 1116-1120.
20. Patil, I., *Visualizations with statistical details: The'ggstatsplot'approach.* Journal of Open Source Software, 2021. **6**(61): p. 3167.
21. Goes van Naters, W. and C. Mucignat-Caretta, *Frontiers in Neuroscience Drosophila Pheromones: From Reception to Perception.* Neurobiology of Chemical Communication. Boca Raton (FL): CRC Press/Taylor & Francis (c), 2014.
22. Brown, D., et al., *Subfamily hmms in functional genomics.* Pac Symp Biocomput, 2005: p. 322-33.
23. Srivastava, P.K., et al., *HMM-ModE--improved classification using profile hidden Markov models by optimising the discrimination threshold and modifying emission probabilities with negative training sequences.* BMC Bioinformatics, 2007. **8**: p. 104.
24. Roberts, M.S., et al., *Treatment of Autosomal Dominant Hypocalcemia Type 1 With the Calcilytic NPSP795 (SHP635).* J Bone Miner Res, 2019. **34**(9): p. 1609-1618.
25. Pruitt, K.D., T. Tatusova, and D.R. Maglott, *NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins.* Nucleic Acids Res, 2007. **35**(Database issue): p. D61-5.
26. Gorvin, C.M., *Calcium-sensing receptor signaling—How human disease informs biology.* Current opinion in endocrine and metabolic research, 2021. **16**: p. 10-18.
27. Bradley, D. and P. Beltrao, *Evolution of protein kinase substrate recognition at the active site.* PLoS biology, 2019. **17**(6): p. e3000341.
28. Chun, L., W.H. Zhang, and J.F. Liu, *Structure and ligand recognition of class C GPCRs.* Acta Pharmacol Sin, 2012. **33**(3): p. 312-23.
29. Pi, M., S.K. Nishimoto, and L.D. Quarles, *GPRC6A: Jack of all metabolism (or master of none).* Mol Metab, 2017. **6**(2): p. 185-193.
30. Nango, E., et al., *Taste substance binding elicits conformational change of taste receptor T1r heterodimer extracellular domains.* Sci Rep, 2016. **6**: p. 25745.
31. Kumari, R., C. Castillo, and A. Francesconi, *Agonist-dependent signaling by group I metabotropic glutamate receptors is regulated by association with lipid domains.* J Biol Chem, 2013. **288**(44): p. 32004-19.
32. Nuemket, N., et al., *Structural basis for perception of diverse chemical substances by T1r taste receptors.* Nat Commun, 2017. **8**: p. 15530.

33. Chen, T. and C. Guestrin. *Xgboost: A scalable tree boosting system*. in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 2016.
34. Chou, K.J., et al., *A new missense mutation of calcium sensing receptor with isoleucine replaced by serine at codon 857 leading to type V Bartter syndrome.* Exp Cell Res, 2022. **414**(1): p. 113080.
35. Moon, J.E., et al., *A cell function study on calcium regulation of a novel calcium-sensing receptor mutation (p.Tyr825Phe).* Ann Pediatr Endocrinol Metab, 2021. **26**(1): p. 24-30.
36. Palmieri, S., et al., *Case Report: Unusual Presentations of Loss-of-Function Mutations of the Calcium-Sensing Receptor.* Front Med (Lausanne), 2021. **8**: p. 809067.
37. Li, N., et al., *A novel homozygous mutation of the calcium-sensing receptor gene associated with apparent autosomal recessive inheritance of familial hypocalciuric hypercalcemia.* Chin Med J (Engl), 2021. **134**(15): p. 1869-1871.
38. Hao, Y., et al., *Radiofrequency Ablation of Parathyroid Glands to Treat a Patient With Hypercalcemia Caused by a Novel Inactivating Mutation in CaSR.* Front Endocrinol (Lausanne), 2021. **12**: p. 743517.
39. Tsuji, T., et al., *Autosomal Dominant Hypocalcemia With Atypical Urine Findings Accompanied by Novel CaSR Gene Mutation and VitD Deficiency.* J Endocr Soc, 2021. **5**(3): p. bvaa190.
40. Dharmaraj, P., et al., *Neonatal Hypocalcemic Seizures in Offspring of a Mother With Familial Hypocalciuric Hypercalcemia Type 1 (FHH1).* J Clin Endocrinol Metab, 2020. **105**(5).
41. Gomes, V., et al., *Autosomal dominant hypocalcaemia: identification of two novel variants of CASR gene.* BMJ Case Rep, 2020. **13**(6).
42. Magno, A.L., et al., *Functional Analysis of Calcium-Sensing Receptor Variants Identified in Families Provisionally Diagnosed with Familial Hypocalciuric Hypercalcaemia.* Calcif Tissue Int, 2020. **107**(3): p. 230-239.
43. Majumdar, S.K., et al., *A Novel Variant in the Calcium-Sensing Receptor Associated with Familial Hypocalciuric Hypercalcemia and Low-to-Normal PTH.* Case Rep Endocrinol, 2020. **2020**: p. 8752610.
44. Zajickova, K., et al., *Familial hypocalciuric hypercalcemia in an index male: grey zones of the differential diagnosis from primary hyperparathyroidism in a 13-year clinical follow up.* Physiol Res, 2020. **69**(Suppl 2): p. S321-S328.

45. Sagi, S.V., et al., *A novel CASR variant in a family with familial hypocalciuric hypercalcaemia and primary hyperparathyroidism.* Endocrinol Diabetes Metab Case Rep, 2020. **2020**.

46. Dong, Q., et al., *[Clinical and genetic analysis of a child with neonatal severe parathyroidism].* Zhonghua Yi Xue Yi Chuan Xue Za Zhi, 2020. **37**(11): p. 1247-1249.

47. Hawkes, C.P., D.I. Shulman, and M.A. Levine, *Recombinant human parathyroid hormone (1-84) is effective in CASR-associated hypoparathyroidism.* Eur J Endocrinol, 2020. **183**(6): p. K13-K21.

48. Wejaphikul, K., et al., *Subtotal parathyroidectomy successfully controls calcium levels of patients with neonatal severe hyperparathyroidism carrying a novel CASR mutation.* Hormone Research in Paediatrics, 2023: p. 1-7.

49. Courtney, A., et al., *Familial hypocalciuric hypercalcaemia type 1 caused by a novel heterozygous missense variant in the CaSR gene, p (His41Arg): two case reports.* BMC Endocrine Disorders, 2022. **22**(1): p. 324.

50. Bletsis, P., et al., *A Novel missense CASR gene sequence variation resulting in familial hypocalciuric hypercalcemia.* AACE Clinical Case Reports, 2022. **8**(5): p. 194-198.

51. Wu, Y., et al., *Autosomal dominant hypocalcemia with a novel CASR mutation: a case study and literature review.* Journal of International Medical Research, 2022. **50**(7): p. 03000605221110489.

52. Coughlan, A., F. Khan, and M. Brassill, *A Novel Genetic Variant Resulting in Familial Hypocalciuric Hypercalcaemia.* Irish Medical Journal, 2022. **115**(2): p. 545-545.

53. Mullin, B.H., et al., *Functional assessment of calcium-sensing receptor variants confirms familial hypocalciuric hypercalcemia.* Journal of the Endocrine Society, 2022. **6**(5): p. bvac025.

54. Jones, E.M., et al., *Structural and functional characterization of G protein–coupled receptors with deep mutational scanning.* Elife, 2020. **9**: p. e54895.

55. Alirezaie, N., et al., *ClinPred: prediction tool to identify disease-relevant nonsynonymous single-nucleotide variants.* The American Journal of Human Genetics, 2018. **103**(4): p. 474-483.

56. Ioannidis, N.M., et al., *REVEL: an ensemble method for predicting the pathogenicity of rare missense variants.* The American Journal of Human Genetics, 2016. **99**(4): p. 877-885.

57. Rentzsch, P., et al., *CADD: predicting the deleteriousness of variants throughout the human genome.* Nucleic acids research, 2019. **47**(D1): p. D886-D894.

58. Rogers, M.F., et al., *FATHMM-XF: accurate prediction of pathogenic point mutations via extended features.* Bioinformatics, 2018. **34**(3): p. 511-513.
59. Eddy, S.R., *Accelerated profile HMM searches.* PLoS computational biology, 2011. **7**(10): p. e1002195.
60. Finn, R.D., et al., *The Pfam protein families database: towards a more sustainable future.* Nucleic acids research, 2016. **44**(D1): p. D279-D285.
61. *UniProt: the universal protein knowledgebase.* Nucleic acids research, 2017. **45**(D1): p. D158-D169.
62. Altschul, S.F., et al., *Basic local alignment search tool.* Journal of molecular biology, 1990. **215**(3): p. 403-410.
63. Yamada, K.D., K. Tomii, and K. Katoh, *Application of the MAFFT sequence alignment program to large data—reexamination of the usefulness of chained guide trees.* Bioinformatics, 2016. **32**(21): p. 3246-3251.
64. Stamatakis, A., *RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies.* Bioinformatics, 2014. **30**(9): p. 1312-1313.
65. Huerta-Cepas, J., F. Serra, and P. Bork, *ETE 3: reconstruction, analysis, and visualization of phylogenomic data.* Molecular biology and evolution, 2016. **33**(6): p. 1635-1638.
66. Capella-Gutierrez, S., J.M. Silla-Martinez, and T. Gabaldon, *trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses.* Bioinformatics, 2009. **25**(15): p. 1972-3.
67. Kozlov, A.M., et al., *RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference.* Bioinformatics, 2019. **35**(21): p. 4453-4455.
68. Nguyen, L.-T., et al., *IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies.* Molecular biology and evolution, 2015. **32**(1): p. 268-274.
69. Kalyaanamoorthy, S., et al., *ModelFinder: fast model selection for accurate phylogenetic estimates.* Nature methods, 2017. **14**(6): p. 587-589.
70. Hoang, D.T., et al., *UFBoot2: improving the ultrafast bootstrap approximation.* Molecular biology and evolution, 2018. **35**(2): p. 518-522.
71. Waterhouse, A., et al., *SWISS-MODEL: homology modelling of protein structures and complexes.* Nucleic acids research, 2018. **46**(W1): p. W296-W303.
72. Pettersen, E.F., et al., *UCSF Chimera--a visualization system for exploratory research and analysis.* J Comput Chem, 2004. **25**(13): p. 1605-12.

73. Chen, T., et al., *Xgboost: extreme gradient boosting. R package version 04-2.* OS Independent, 2015.
74. Pedregosa, F., et al., *Scikit-learn: Machine learning in Python.* the Journal of machine Learning research, 2011. **12**: p. 2825-2830.
75. Zimmerman, J., N. Eliezer, and R. Simha, *The characterization of amino acid sequences in proteins by statistical methods.* Journal of theoretical biology, 1968. **21**(2): p. 170-201.
76. Bhaskaran, R. and P. Ponnuswamy, *Positional flexibilities of amino acid residues in globular proteins.* International Journal of Peptide and Protein Research, 1988. **32**(4): p. 241-255.
77. Barker, W.C., L.K. Ketcham, and M.O. Dayhoff, *A comprehensive examination of protein sequences for evidence of internal gene duplication.* Journal of Molecular Evolution, 1978. **10**: p. 265-281.
78. Rose, G.D., et al., *Hydrophobicity of amino acid residues in globular proteins.* Science, 1985. **229**(4716): p. 834-838.
79. Kyte, J. and R.F. Doolittle, *A simple method for displaying the hydropathic character of a protein.* Journal of molecular biology, 1982. **157**(1): p. 105-132.
80. Grantham, R., *Amino acid difference formula to help explain protein evolution.* science, 1974. **185**(4154): p. 862-864.
81. Walker, J.M., *The proteomics protocols handbook*. 2005: Springer.
82. Lundberg, S.M. and S.-I. Lee, *A unified approach to interpreting model predictions.* Advances in neural information processing systems, 2017. **30**.
83. Mayer, M. *shapviz: SHAP Visualizations. R package version 0.9.0*. 2023; Available from: https://github.com/ModelOriented/shapviz.

833

834

835