PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Evolutionary history of calcium-sensing receptors unveils hyper/hypocalcemia-causing mutations

Aylin Bircan[1], **Nurdan Kuru**[1], **Onur Dereli**[1], **Berkay Selçuk**[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

¤ Current address: Department of Microbiology, The Ohio State University, Columbus, Ohio, United States of America
* oadebali@sabanciuniv.edu

OPEN ACCESS

Citation: Bircan A, Kuru N, Dereli O, Selçuk B, Adebali O (2024) Evolutionary history of calcium-sensing receptors unveils hyper/hypocalcemia-causing mutations. PLoS Comput Biol 20(11): e1012591. https://doi.org/10.1371/journal.pcbi.1012591

Editor: Jens-Uwe Ulrich, Robert Koch Institute: Robert Koch Institut, GERMANY

Received: February 16, 2024

Accepted: October 23, 2024

Published: November 12, 2024

Data Availability Statement: All code written in support of this publication is publicly available at https://github.com/CompGenomeLab/CaSR.

Funding: This study was supported by EMBO Installation Grant no:4163 funded by TÜBİTAK (to OA). OA is supported by the BAGEP program of the Science Academy - Türkiye, and the TÜBA-GEBİP program of the Turkish Academy of Sciences. BS is supported by EMBO Installation Grant no:4163 funded by TÜBİTAK. The funders had no role in

Abstract

Despite advancements in understanding the structure and functions of the Calcium Sensing Receptor (CaSR), gaps persist in our knowledge of the specific functions of its residues. 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, emphasizing the critical role of residues with high SDP scores 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, providing valuable insights into the regulation of calcium homeostasis and its connections to associated disorders.

Author summary

CaSR plays a crucial role in maintaining calcium balance in the body. Our study investigates the evolutionary history of CaSR to better understand its function and the impact of mutations. By using phylogenetic techniques, we identified key residues that are critical for CaSR's function. These residues are highly conserved, indicating their importance in receptor activation and calcium regulation.

We applied machine learning methods to differentiate between mutations that cause either a gain or loss of function in CaSR, leading to disorders like hypercalcemia and hypocalcemia. Our analysis highlights specific regions of the receptor that are essential for its activity, providing insights into how mutations can disrupt calcium homeostasis.

study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Introduction

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

Class C GPCRs are obligate 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 sites at the interdomain cleft of LB1-LB2 [3–5,11–13] and multiple Ca2+ binding sites on the VFT domain are shown in the literature. [3–5]. While Ca^{2+} serves as the primary agonist for the CaSR, L-amino acids enhance receptor activation in conjunction with Ca^{2+}. However, it's important to note that L-amino acids alone are insufficient to activate the receptor independently [14]. Even though Ca^{2+} 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 Ca^{2+} homeostasis diseases in humans. More than 400 germline loss/gain-of-function (i.e., LoF and GoF, respectively) mutations cause hypercalcaemic disorders, neonatal severe hyperparathyroidism (NSHPT), 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]. Gaining insight into the function of individual residues within the receptor structure and their involvement in activation mechanisms has the potential to enhance our understanding of the probability of variant pathogenicity and the signaling processes of the CaSR. The examination of receptors within a family and across different families allows for the identification of the specific function of each residue in a receptor. However, the comprehensive understanding of the structure and activation mechanisms of several families within the class C GPCRs remains elusive. This is particularly true for the G-protein coupled receptor family C group 6 member A (GPRC6A) and the type 1 taste receptors (TAS1Rs; specifically members 1, 2, and 3), which are the most closely related subfamilies to the CaSR.

While all subfamily receptors of class C GPCRs share common domains and structural features, the details of responding to different ligands and activating signaling pathways may differ even between closely related receptors [6]. Gene duplication is the main mechanism that generates new protein functions across GPCRs [15]. Protein families are evolved by speciation events following gene duplication [16,17]; 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 (specifically, outparalogs that have diverged after a speciation event) and distant homologs may help infer the specific roles of a single residue in protein function. Because the evolutionary pressure on paralogs and close orthologs is not the same, substitutions allowed in paralogs may not be acceptable in close orthologs. While orthologs are derived from a single ancestral gene in the

last common ancestor of the compared species, paralogs are arised from gene duplication event and generally, paralogs perform biologically distinct functions [18]. Thus, using functionally identical orthologs in sequence comparisons is crucial to inferring 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 orthologous sequences in each subfamily by building phylogenetic trees and manually curating the duplications/speciations on the tree to obtain all functionally equivalent orthologs within each subfamily. To obtain members of a subfamily without requiring computationally expensive phylogenetic tree building and manual curation steps, we generated highly sensitive subfamily-specific profile hidden Markov models (HMMs) by using the functionally equivalent orthologs we determined using phylogenetic tree analysis. We calculated a specificity score for each residue in a subfamily by calculating scores based on a modified version of the PHACT algorithm [19] 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 every potential substitution in CaSR by using the gradient boosting trees machine learning approach. Lastly, we investigated how our predictions relate to the activation mechanism of CaSR.

Results

Functionally equivalent orthologs and evolutionary history of CaSR in Class C

To reveal the evolutionary constraints on protein families, we developed a strategy to precisely define a protein subfamily. A precise subfamily definition can be achieved by revealing the evolutionary history of the superfamily. The 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 the paralogous and orthologous protein sequences. We further analyzed the phylogenetic trees to classify the orthologous sequences that are likely equivalent in function. A subfamily is defined by the human receptor and its orthologs. Within class C GPCRs, there are twenty-two distinct subfamilies: CaSR, GPRC6A, three taste receptors (TAS1R1-3), eight metabotropic glutamate receptors (mGluR1-8), GPRC5A-D, GABBR1, GABBR2, GPR156, GPR158, and GPR179. 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 the class C GPCR family, we performed a hmmsearch using 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 did not provide subfamily annotations for the 22 subfamilies in class C GPCRs. To select canonical isoforms, we performed a profile hmmscan of the PfamA profile against Class C GPCR. To generate a general HMM profile for each subfamily, we first applied a BLAST search using each human class C GPCR as a query [20]. For each subject, we blasted them against the human proteome and retrieved the bidirectional best hits (Core subfamily assignment, **Fig** 1).

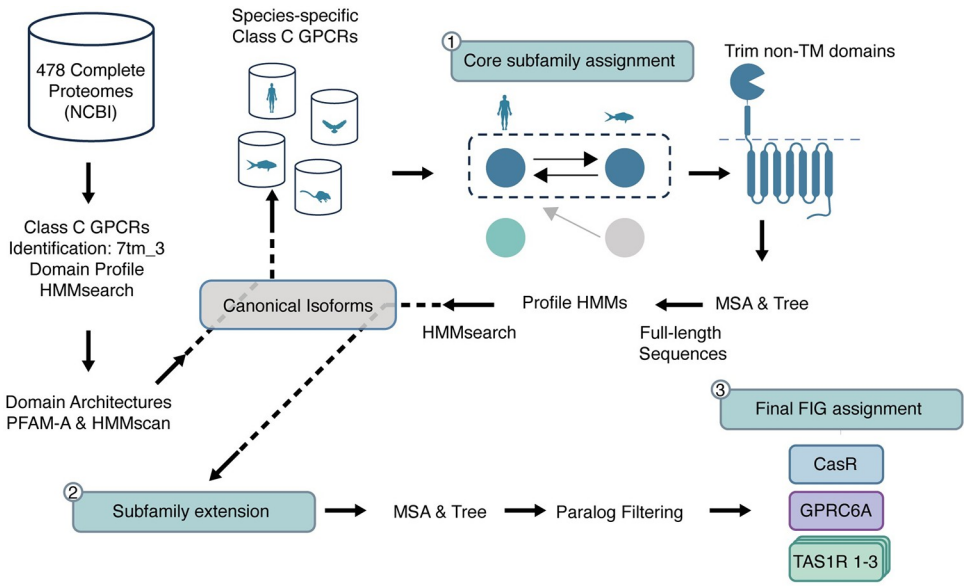


Fig 1. Summary of the Methodological Framework. 478 complete proteomes were retrieved from the NCBI database. Each sequence was searched by hmmsearch against the Pfam 7tm_3 domain profile to retrieve all class C GPCRs. Domain architectures of class C GPCRs were determined by hmmscan against Pfam. A profile was created 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.

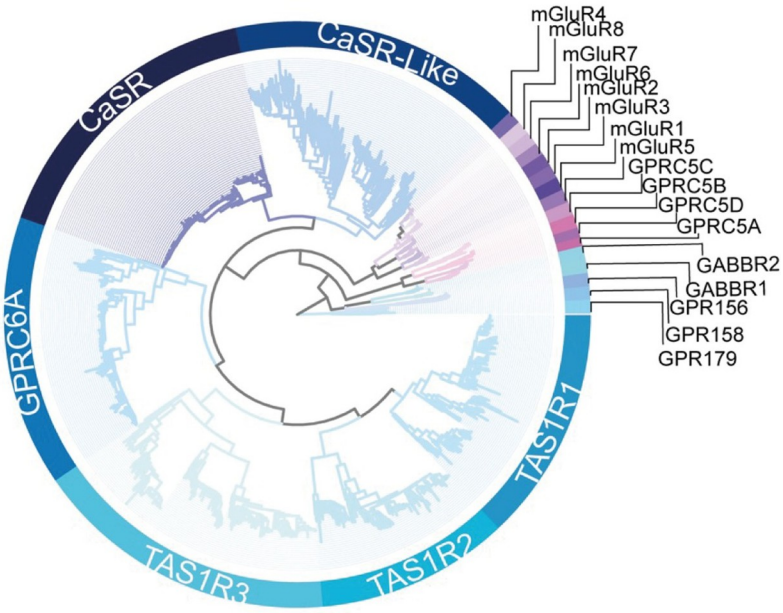
https://doi.org/10.1371/journal.pcbi.1012591.g001

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 inparalogous sequences to obtain functionally identical groups.

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 [21]. In species that had multiple proteins assigned to the CaSR subfamily, we constructed an ML 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 (a group of organisms that includes a common ancestor and all of its descendants) diverged from CaSR. As a result, we defined this diverged clade as a new subfamily named CaSR-likes. In this subfamily, there are diverse sequences including vomeronasal, olfactory receptors that are 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 an ML tree (Fig 2A). 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 2A). We used 1000 transfer bootstrap to ensure statistical reliability of the phylogenetic tree. We rooted the ML tree using the GPR158 and GPR179 as outgroups. There are 22 different class C GPCRs in humans. 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

A


CaSR
CaSR-Like
GPRC6A
TAS1R1
TAS1R2
TAS1R3
mGluR4
mGluR8
mGluR7
mGluR6
mGluR2
mGluR3
mGluR1
mGluR5
GPRC5C
GPRC5B
GPRC5D
GPRC5A
GABBR2
GABBR1
GPR156
GPR158
GPR179

B

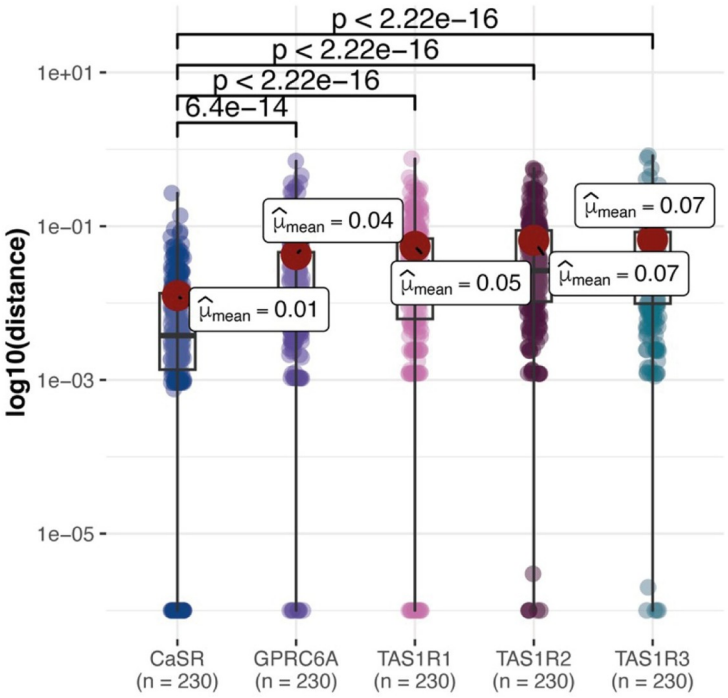

p < 2.22e−16
p < 2.22e−16
p < 2.22e−16
6.4e−14
1e+01
1e−01
1e−03
1e−05
log10(distance)
μ̂mean = 0.01
μ̂mean = 0.04
μ̂mean = 0.05
μ̂mean = 0.07
μ̂mean = 0.07
CaSR (n = 230)
GPRC6A (n = 230)
TAS1R1 (n = 230)
TAS1R2 (n = 230)
TAS1R3 (n = 230)

Fig 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 into this representative species tree. **(B)** Branch lengths from leaf to root of the common species that exist in all CaSR, GPRC6A and TAS1Rs are taken from the subfamily trees. Welch's t-Test using the ggstatsplot package results are shown on the graph.

https://doi.org/10.1371/journal.pcbi.1012591.g002

diverged from other receptors consistent with previous trees [22] and had a 0.95 transfer bootstrap expectation (TBE) value. γ-aminobutyric acid-B receptors (GABBR1 and GABBR2) formed another clade diverging from GPR156 with 0.97 TBE. γ-aminobutyric acid-B receptors evolved earliest and have a common ancestor with the highest taxonomic rank (33213––Bilateria) compared to other subfamilies. The CaSR group (CaSR, CaSR-likes, GPRC6A and taste receptors) was diverged from metabotropic glutamate receptors (mGluR1-8) and RAIG receptors (GPRC5A, GPRC5B, GPRC5C, GPRC5D) with 1 and 0.98 TBE values, respectively. Within the CaSR group, clade CaSRs and CaSR-likes were diverged from GPRC6A and taste receptors with 1 TBE. Except for 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 evolved from 117571-Euteleostomi. In a phylogenetic tree, the length of branches signifies genetic changes over time, where longer branches indicate greater divergence. Employing Welch's t-Test in our statistical analysis, we measured these branch lengths to gain insights into the evolutionary conservation of receptor subfamilies, revealing genetic similarities and differences. Our comparison analysis of branch lengths [22] among common species in CaSR, GPRC6A, and taste receptors highlights a significant conservation trend within the CaSR subfamily compared to its closest counterparts (Fig 2B).

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 matching the CaSR profile. However, these matches include type 2 vomeronasal receptors (V2R) and V2R-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 [23]. Since these receptors do not have functional orthologs in mammals, they are likely to be functionally diverged, and it is crucial to separate them from functionally-equivalent CaSRs.

Precision in subfamily identification: Constructing subfamily-specific profile hmms for class c gpcrs based on phylogenetic analysis

In the class C GPCR family, gene duplication events give rise to new specificity, and each duplicated gene with a new function evolves through further speciation events, producing a set of orthologous sequences [16,17]. Each subfamily of class C GPCRs shares a relatively conserved membrane-spanning region but also exhibits a degree of variability that underlies 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 [16]. Conservation analysis from multiple sequence alignments (MSAs) can be used to find residues that are conserved in all subfamilies through evolution as well as specificity-

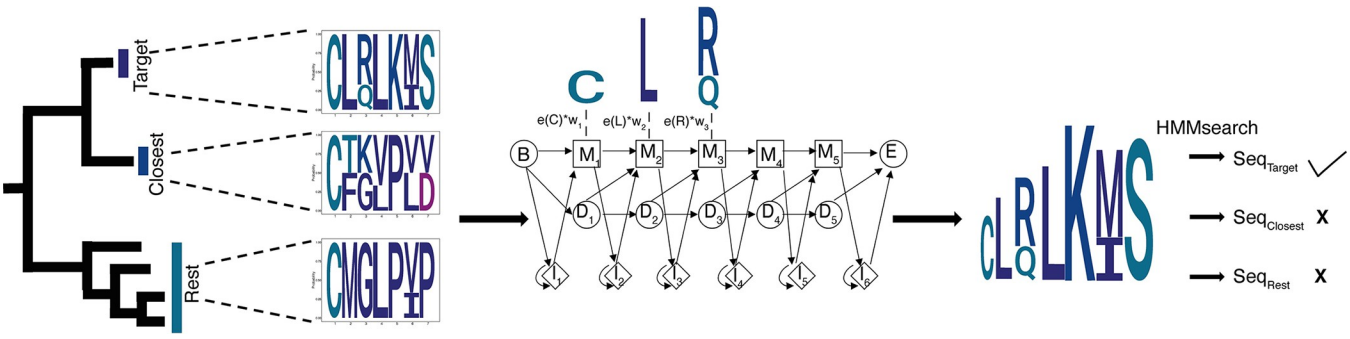


Fig 3. Subfamily Specific HMM Models. Based on the phylogenetic tree, the target, the closest, and the rest groups were determined. Initial HMMs were built without using priors. Representative amino acids in each group are selected, and their scores are calculated. According to groups, representative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

https://doi.org/10.1371/journal.pcbi.1012591.g003

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. Therefore, 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 does not contain enough information to differentiate subfamilies further. Moreover, excessive gene duplication events, as seen in the CaSR-like clade, require precise phylogenetic analysis to differentiate between 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 [24,25]. 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 Fig 3.

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 downloaded complete proteomes from UNIPROT [26] and we tested our subfamily-specific profile HMMs' performance on independent sequences retrieved from UNIPROT [26]. These sequences were from different species and they were not used in calculating the position weights. We assigned sequences to their corresponding subfamilies by following the same steps as the NCBI dataset [27] used to build these models. We selected new taxa that were not in the NBCI dataset to test the performance of our profiles, and our subfamily specific profile HMMs correctly identified all members of a subfamily while avoiding hits to proteins from other subfamilies (Table 1).

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

https://doi.org/10.1371/journal.pcbi.1012591.t001

Uncovering molecular distinctions: revealing specificity-determining residues in casr and its closest relatives GPRC6A, TAS1R1-3

CaSR is distinguished from other subfamilies of class C GPCRs by its oversensitivity to substitutions that can result in either GoF or LoF mutations. This sensitivity is due to its critical role in maintaining systemic calcium homeostasis and its high responsiveness to very slight changes in extracellular Ca^{2+} concentrations [28]. As CaSR is the most conserved among CaSR-likes, GPRC6A, and TAS1Rs, it is reasonable to anticipate that certain positions may experience a relaxation of purifying selection in CaSR-likes, GPCR6A, or TAS1Rs. However, CaSRs exhibit a greater number of positions that are subject to negative Darwinian selection compared to other sub-families. Conversely, at some positions, the same amino acid may retain functional significance in both subfamilies, and at others, a position remains important in each subfamily, but different amino acids are favored in each duplicate.

Specificity-determining residues that are conserved in subfamily, but differ from its sister clade can be predicted by directly comparing ancestral family sequences and calculating their divergence scores [29]. However, using MSAs alone does not account for the number of substitution events. For example, a single substitution event in the common ancestor of the bony fish clade of the CaSR subfamily can be inherited into multiple descendants' sequences. Assessing this single substitution event as it repeats in each sequence independently results in overcounting of these changes. Due to this mistake and overcounting the effect of one single mutation repeatedly, 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 CaSR-likes subfamily, it indicates that the amino acid has a CaSR-specific role. We expect the SDP score of such an amino acid to be high compared to others. 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.

To consider these two important factors, we designed an approach to identify and prioritize residues by specificity, which differentiate a subfamily from others in the CaSR group (CaSR, GPRC6A and TAS1Rs) (Fig 4A). Our approach is based on the idea presented in the functionally divergent residues method *[29]* along with adaption of the PHACT method [19]. We identified specificity-determining residues by tracking substitution events from the root of the tree, counting independent events within subfamily clades, and comparing probabilities across subfamily clades. Our approach, detailed in materials and methods (Algorithm 3), prioritizes residues that are variably conserved in sister subfamily nodes but highly conserved in the target subfamily node, signifying their specificity to the target subfamily.

We calculated specificity scores for each CaSR, GPRC6A and TAS1Rs. We have found that CaSR has residues with high SDP scores on different domains (Fig 4B). Cytoplasmic domain and extracellular loop include less specific residues compared to other domains. Specificity score distributions show that CaSR has more specific residues compared to other subfamilies (Fig 4C). On the VFT domain, specific residues are clustered in different regions (Fig 5). We found a cluster of specific residues on the interdomain cleft between LB1-LB2 which is the L-amino acid binding site in other class C GPCRs [3]. It suggests that this region is the primary Ca^{2+} binding site in CaSR, consistent with [14]. We found two different clusters of specific

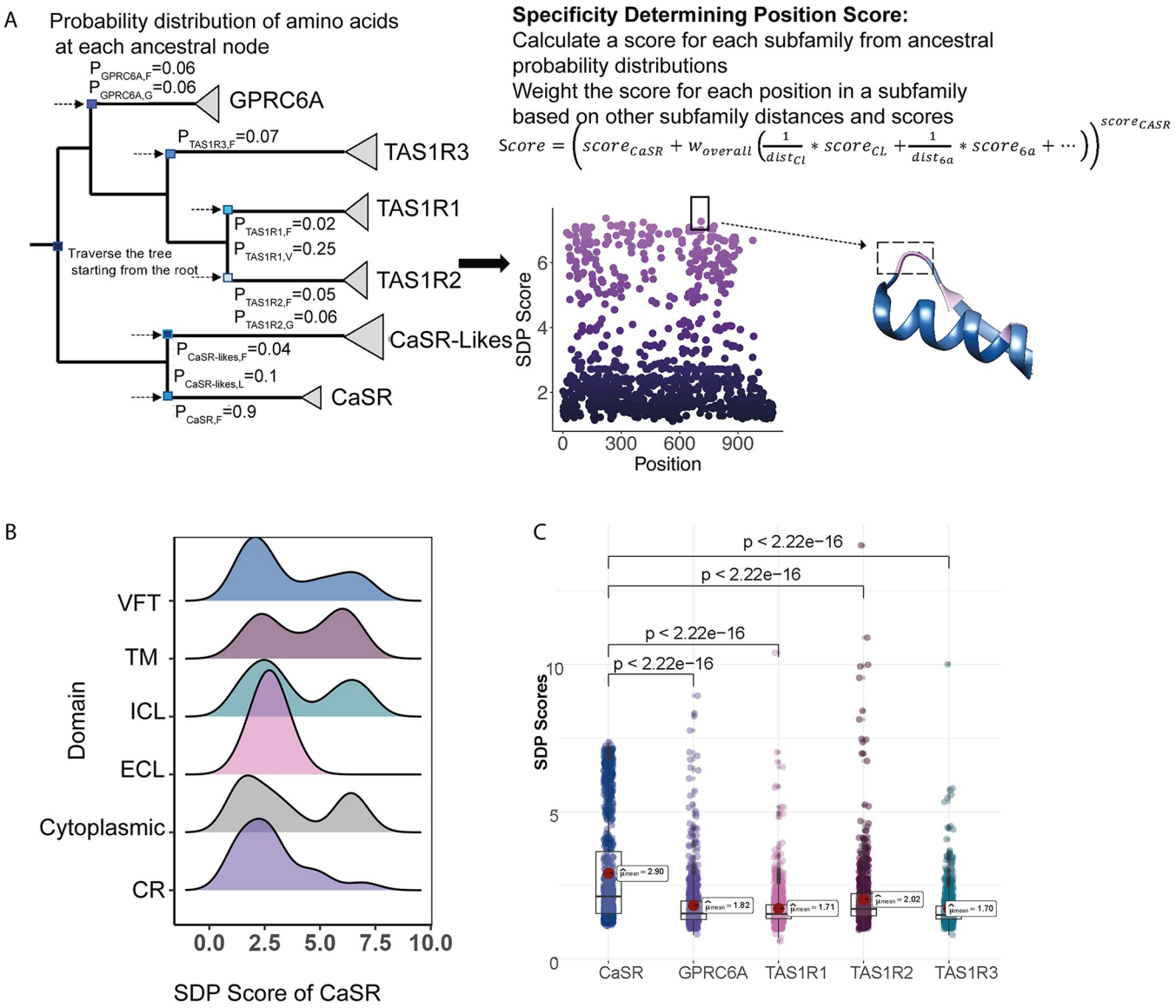


Fig 4. Specificity Determining Position Scores. (A) The calculation of SDP scores uses the phylogenetic tree and the probability distribution of amino acids at each ancestral node. **(B)** The SDP score distributions of CaSR among different domains and the SDP score distributions of each subfamily are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

https://doi.org/10.1371/journal.pcbi.1012591.g004

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 Ca^{2+} ions bind [3–5]. Interaction between LB2 subunits is 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 in these regions are associated with LoF and GoF mutations [2]. Other specific residues are found on the CR, extracellular loop 2 (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

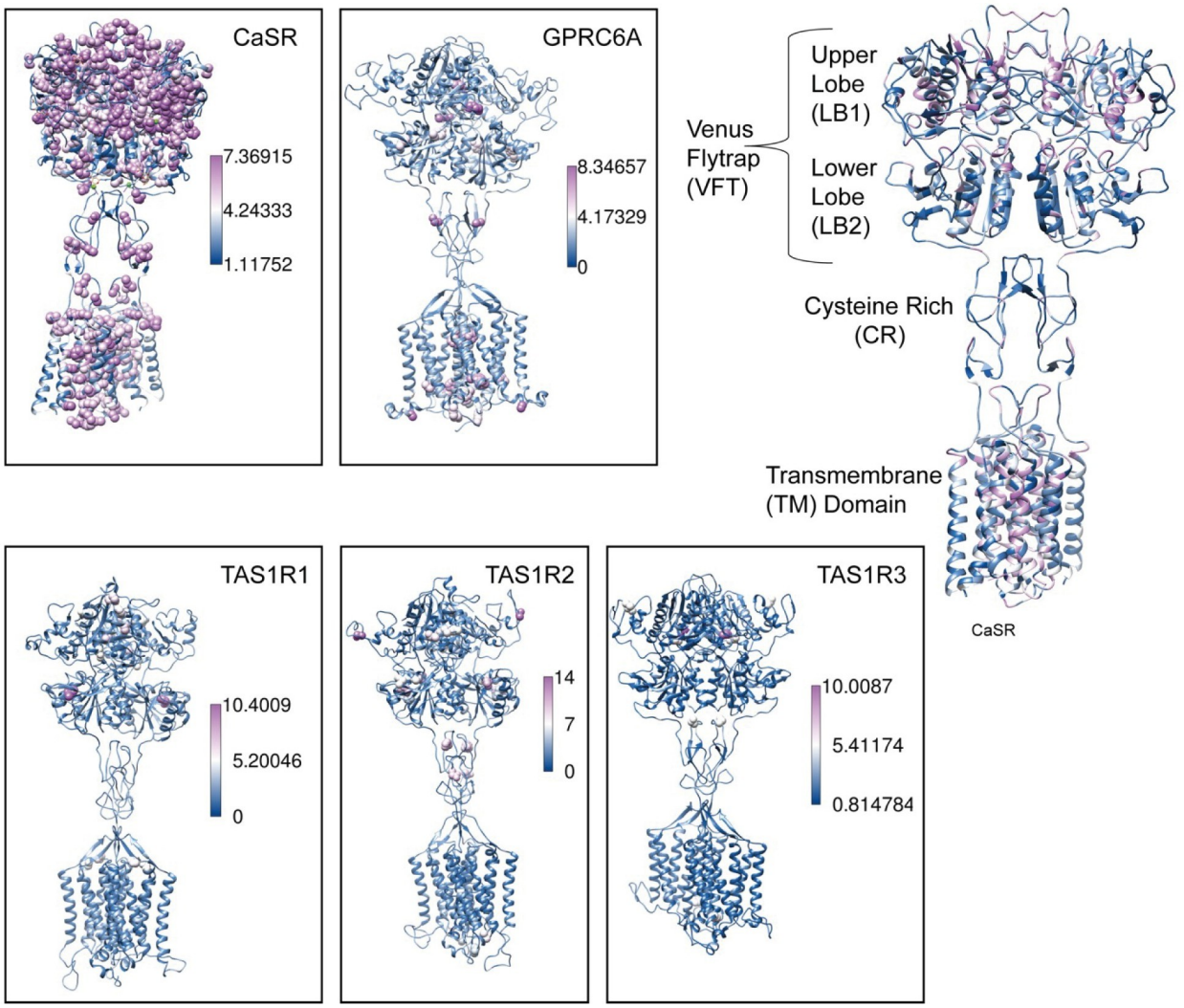


Fig 5. Specific conserved residues mapped onto structures. The cryo-EM structure of human CaSR bound with $Ca^{2+}$and L-Trp (PDB:7DTV) and homology models of GPRC6A, TAS1R1, TAS1R2 and TAS1R3 are colored based on SDP scores. Residues with a high SDP score (above 5.0) are shown as spheres. Domains on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

https://doi.org/10.1371/journal.pcbi.1012591.g005

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 disrupt CaSR activity; however, Δ758–759 mutant was expressed at the cell surface with comparable levels to that of wild type (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 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 the TM domain are likely play a role in the regulation of conformational changes observed during activation upon

ligand binding and inactive states. Residues inside the dimerization interface and interacting with the ligand are quite sensitive substitutions because they can induce malfunctions in the receptor easily. On the other hand, GPRC6A and taste receptors are more tolerant to substitutions, and they are not very specialized respond to a single ion. GPRC6A and taste receptors are activated by a broad spectrum of ligands [30,31]. 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 [31]. Taste receptors bind to different ligands, including sugar, L- and D-amino acids, sweet proteins, and artificial sweeteners [32].

On the TM region, we also find the 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 the mGluR2 receptor [33]. Phylogenetic analysis shows that TAS1R3 evolved earliest (7776 Gnathostomata) among TAS1Rs, TAS1R1and TAS1R2 subfamilies have a common ancestor 117571 Euteleostomi. TAS1R3 forms heterodimers with TAS1R1 and TAS1R2 [30,32,34]. Interactions between the cytosolic terminus of the extracellular CRD is needed for TAS1R3 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, TAS1R3 loses the corresponding function [34]. Our analysis showed that TAS1R1 has specific residues on the LB1, LB2 and extracellular loop regions. Also, TAS1R2 has specific residues on the 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 the 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 the same strict selective pressure as CaSR.

Gradient boosting trees machine learning approach to predict the mutation types in CaSR

Due to the high conservation of the CaSR subfamily, substitutions on the receptor may lead to varied outcomes, including disruption of receptor function and the potential for either GoF or LoF mutations. GoF mutations typically enhance CaSR activity and lead to an increase in the sensitivity to extracellular calcium. LoF mutations reduce or eliminate the function of CaSR. However, predicting the functional consequences of a substitution is challenging. Evolutionary conservation of a residue among subfamilies might reflect the common structural constraints, but it does not distinguish between LoF and GoF mutations. In addition, at certain positions, substitution of different amino acids causes either LoF or GoF mutations [23]. We hypothesized that "activating" mutations are more likely to be tolerated in neighboring clades such as GPRC6A and TAS1Rs, but not in CaSR. Generally, LoF (inactivating) mutations are not tolerated in the larger clade of these receptor subfamilies. To test this hypothesis and to determine whether we can discriminate between GoF and LoF mutations in CaSR, we analyzed 94 GoF and 243 LoF mutations, categorized based on their clinical outcomes from the literature [2]. We utilized a tree boosting machine learning algorithm, XGBoost [35] which incorporates multiple features such as conservation scores, physico-chemical properties of amino acids, and domain information.

Our algorithm uses sequence-based features, identity scores from MSAs, physico-chemical properties of amino acids, and domain information as input features (Fig 6A). Since we calculated our feature values from the MSAs, we divided our dataset into training, validation and

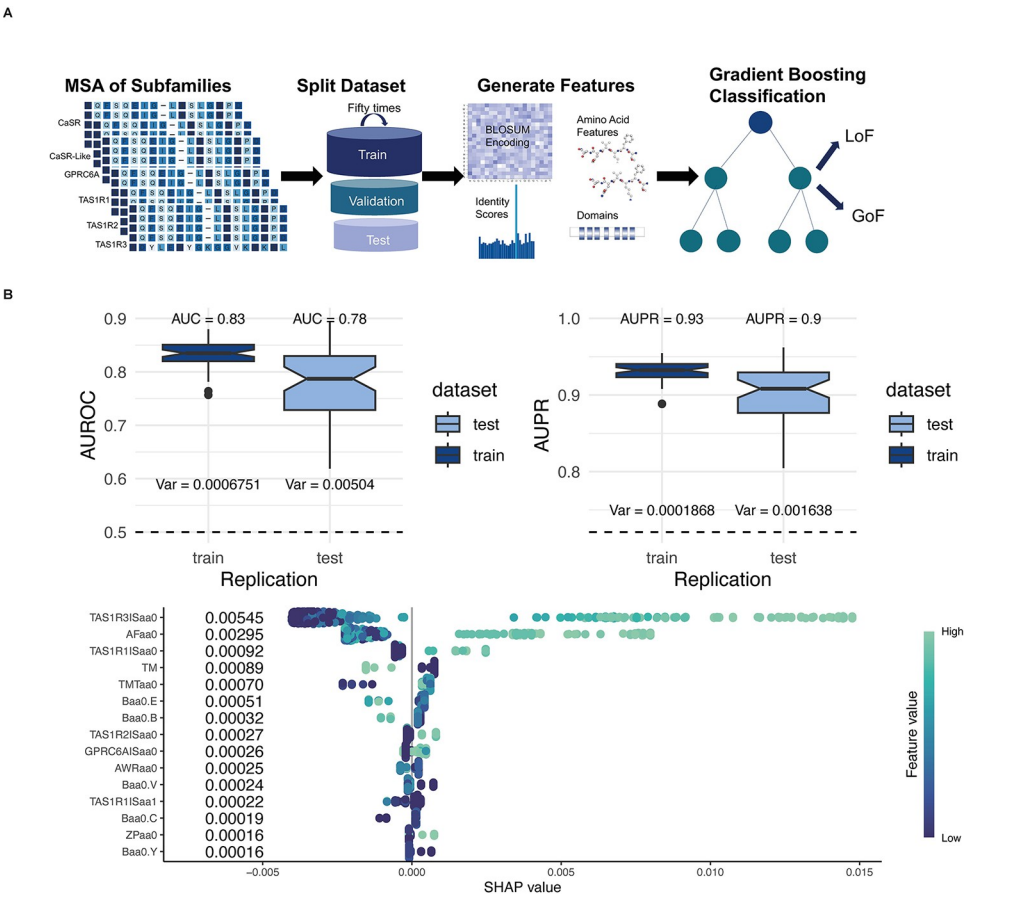


Fig 6. Gradient Boosting Trees Machine Learning Approach to Predict the Mutation Types in CaSR. (A) Model architecture. We took 94 GoF and 243 LoF mutations from the literature. We divided subfamily alignments and mutations randomly as 80% training and the remaining 20% test data before creating feature 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 the LoF to GoF ratio almost the same in the datasets. 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. Further details about these features can be found in materials and methods section.

https://doi.org/10.1371/journal.pcbi.1012591.g006

test datasets before we created feature matrices to prevent information leakage. We performed 50 replications with different random splittings of datasets to obtain a more robust model performance.

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, 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.

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 [36]	hypocalcemia	gain-of-function*
p.Y825F [37]	hypocalcemia	gain-of-function*
p.P393R [38]	hypercalcemia	loss-of-function*
p.C60G [39]	hypercalcemia	loss-of-function*
p.D99N [40]	hypercalcemia	loss-of-function*
p.T186N [41]	hypocalcemia	loss-of-function
p.A840V [26]	hypocalcemia	gain-of-function*
p.S448P [42]	hypercalcemia	loss-of-function*
p.L696V [43]	hypocalcemia	gain-of-function*
p.D433Y [44]	hypercalcemia	loss-of-function*
p.S147L [45]	hypercalcemia	loss-of-function*
p.D398N [46]	hypercalcemia	loss-of-function*
p.K805R [47]	hypercalcemia	gain-of-function
p.C60Y [48]	hypercalcemia	loss-of-function*
p.L606P [49]	hypercalcemia	loss-of-function*
p.H41R [50]	hypercalcemia	gain-of-function
p.A110D [51]	hypercalcemia	gain-of-function
p.I139T [52]	hypocalcemia	gain-of-function*
p.Q164R [53]	hypercalcemia	loss-of-function*
p.T699N [54]	hypercalcemia	gain-of-function
p.R701G [54]	hypercalcemia	loss-of-function*
p.T808P [54]	hypercalcemia	loss-of-function*

https://doi.org/10.1371/journal.pcbi.1012591.t002

AUC and AUPR over all replications were shown in (Fig 6B). Our average AUC values for training and test among 50 replications are 0.83 and 0.78 (Fig 6B). Our average main AUPR values for training and test among 50 replications are 0.93 and 0.9, respectively (Fig 6B). 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. After we reported our algorithm performance, we trained our algorithm with the whole dataset. We tested our algorithm with previously unseen test cases from the literature (Table 2). We provided all predictions (S2 and S3 Tables) and visualized them in the form of a heatmap for every other amino acid at each position until the disordered region (position 892) of the human CaSR (Fig 7A). We mapped known CaSR LoF and GoF mutations on the cryo-EM structure of human CaSR bound with Ca^{2+} and L-Trp (PDB:7DTV)[5] (Fig 7B). 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 Fig 7B. 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 LoF. On the contrary, when an amino acid is conserved exclusively in CaSR, substituting that amino acid is predicted to result in a GoF. New test cases (Y825F, A840V, L696V, I139T) mentioned in Table 2 have been accurately predicted to cause GoF mutations. These specific amino acids are conserved either solely in CaSR or in GPRC6A and CaSR-like receptors, but not across all receptor types. Notably, the A840V substitution leads to a GoF mutation, as

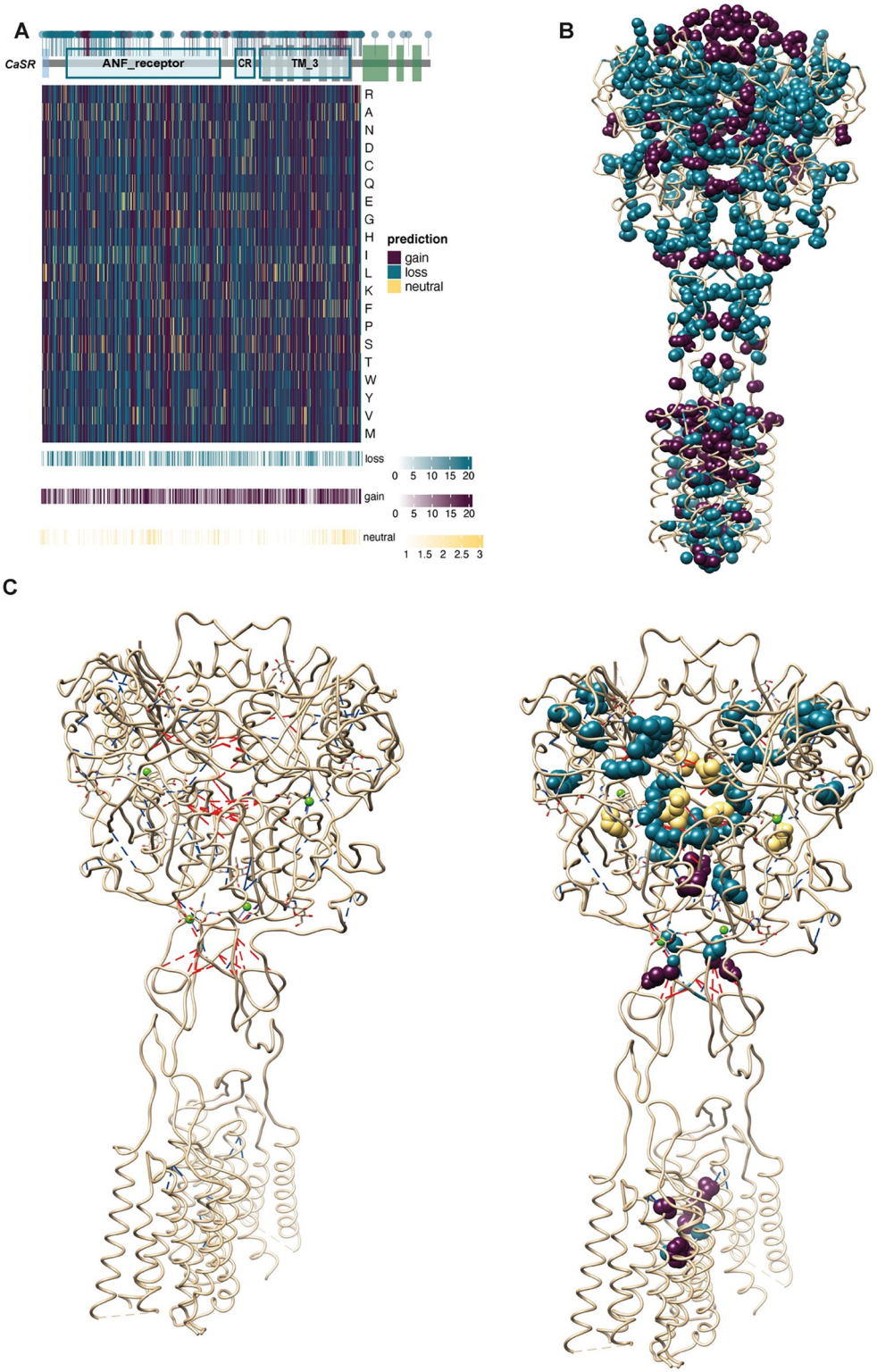


Fig 7. LoF and GoF Mutation Predictions. (A)Visualizing the results of our XGBoost model. The heatmap displays the XGBoost model's predictions for each of the 20 amino acids at every position except disordered regions (892–1078) within the human CaSR. Above the heatmap, the domains of the CaSR are shown. Within these domains, circles represent all known LoF and GoF mutations documented in the literature. Circles denoting GoF mutations are colored purple, while those representing LoF mutations are colored blue. Below the heatmap bar graphs show the number of

GoF, LoF and neutral predictions among the 19 possible substitutions. **(B)** Mutations on the human CaSR structure. LoF- and GoF-associated mutations are shown on the cryo-EM structure of human CaSR bound with Ca^{2+} and L-Trp (PDB:7DTV) as blue and red spheres, respectively. **(C)** Increased residue-residue contacts are shown on the cryo-EM structure of human CaSR bound with Ca^{2+} and L-Trp (PDB:7DTV) on the left. Interdomain and intrasubunit interactions are shown as red and blue lines, respectively on the right. LoF- and GoF-associated mutations among the interacted residues are shown as blue and purple spheres, respectively. Switch residues are shown as yellow spheres.

https://doi.org/10.1371/journal.pcbi.1012591.g007

valine is conserved in GPRC6A. These analyses reinforce our hypothesis that residues conserved in CaSR, but not in other subfamilies, are more likely to induce GoF mutations. Additionally, amino acids conserved within closely related subfamilies might disrupt the functioning of other subfamilies. On the other hand, the new LoF cases shown in Table 2 (D99N, C60G, Q164R, T808P) show that reference amino acids are conserved across subfamilies.

To compare our model's performance, we retrieved 922 missense CaSR mutations from gnomAD v4 [55]. We filtered mutations with known significance. We classified 'pathogenic,' 'likely pathogenic,' and 'pathogenic/likely pathogenic' annotated mutations as pathogenic, while those annotated as 'benign,' 'likely benign,' and 'benign/likely benign' were classified as benign. We categorized mutations predicted by our model and LoGoFunc [56] as 'pathogenic' (including both GoF and LoF mutations) and 'benign' (neutral predictions). We omitted 'ambiguous' predictions from the AlphaMissense [57] tool to ensure clarity in our comparison. We obtained better results compared to others (S1 Fig). We also compared our tool's gain and loss prediction performance with the LoGoFunc tool [56]. We used 22 mutations that had not been previously encountered during our model's development. Our approach demonstrated an accuracy of 0.77 and an F1 score of 0.83, while the LoGoFunc tool exhibited an accuracy of 0.77 and an F1 score of 0.85 (S1 Table).

Another important feature is the structural location of the amino acid. Our findings indicate that if the amino acid is located in the TM domain, a substitution would result in a GoF mutation. It is known that the majority of GoF mutations are located in the TM domain, as shown in Fig 7B. 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.

To gain insight into the activation mechanism of the CaSR and its association with activating and inactivating mutations, we analyzed the complex network of residue-residue connections by comparing active state CaSR structures with inactive state structures. For the monomer activation network, we modified the residue-residue contact score (RRCS) algorithm [58] to detect atom level contact changes in identical amino acids in different states. We made this change for two main reasons. First, class C GPCRs do not exhibit large structural changes during receptor activation [59], like other GPCR classes. Therefore, residue level changes do not provide enough resolution for understanding the mechanism, especially for the TM domain. Second, residue level changes can only highlight certain important positions while investigating atom level contact changes enables us to understand the impact of more residues. For the changes observed within the dimerization interface, we used the original algorithm. To build the networks, we identified significant contact changes observed in two receptor states during activation both in the individual protomer and at dimer interface. (Fig 7C). Significantly, the residue pairs involved in CasR activation were more abundant in critical locations such as the loop of lobe 1 inside the VFT domain, the interdomain cleft, the CR domain, and the TM domains (Fig 7C).

Recent studies have highlighted the role of domain twisting in the CaSR homodimer's activation process, initiated by intersubunit domain contacts and conformational shifts at the interface between the lobe 1 regions (Ling et al., 2021) [5]. In this study, we have noticed a notable augmentation in the level of contact between two pivotal amino acid residues, L125 and Y20, located within the loop region of LB1 domains in the homodimeric structure. Notably, the substitutions of L125F and L125P led to mutants that exhibited a GoF phenotype. In the TM6 domain, we observed increased RRCS between residues in the dimer, with particular attention to A824 and P823. Significantly, residue A824, which exhibits specific conservation, has been associated with GoF mutations in the A824P and A824S [2]

We identified a notable increase in RRCS between residues within the CR domain. Specifically, these interactions have been observed between two subunits of the dimer and involve the following interactions: E556-D578, E556-S580, and E556-K552. Although only residue S580 demonstrated specificity for CaSR, other acidic residues either exhibited tolerance towards replacements with other acidic residues or remained conserved across all subfamilies of CaSR. Nevertheless, it is important to acknowledge that the substitution of E556K resulted in a mutation with enhanced activation. The results indicate that the interactions between these particular residues are of utmost importance in maintaining the receptor's active conformational state. This conclusion is consistent with previous structural investigations that have demonstrated the convergence of LB2, CR, and TM domains in both subunits during the twisting of CaSR, leading to a more condensed conformation in the active state of CaSR (Chen et al., 2021; Ling et al., 2021) [3,5]. In contrast, we have also observed distinct residues linked to LoF mutations. These residues exhibit enhanced interactions within a single subunit, predominantly located in the LB1 and LB2 sections of the VFT domain, specifically in proximity to the Ca^{2+} binding sites (Fig 7C). Residues that induce LoF upon mutation are primarily located within the core of the VFT domain. This implies that any modification in amino acids could potentially induce structural alterations, ultimately resulting in misfolding or disruption of the activation mechanism and consequent LoF.

Discussion

In this study, we showed the evolution of CaSR by developing a methodology for 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 are able to tolerate substitutions at most of the positions, CaSR requires a delicate balance for proper functioning.

The high evolutionary conservation and specificity of CaSR in contrast to the closest subfamilies are reflected in the SDP score analysis. CaSR has specific residues clustered in different regions of the receptor. They are located on Ca^{2+} 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 the correct activation of the CaSR. Ca^{2+} ion binding and interactions between LB2-CR domains and conformational changes in LB1 domain were suggested to be required to activate CaSR [3–5]. Mutational analysis at some positions on the LB1 domain has been shown to reduce the effect of Ca^{2+}-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 a 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 a high SDP score might either cause over or less activation. Deep mutational scanning approaches or new methods that simultaneously profile variant libraries [60] 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 the Extreme Gradient Boosting (XGBoost) method. XGBoost is able to perform well on small datasets by incorporating a 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 GoF 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 LoF and the number of GoF mutations for training and test sets as close as possible. To ensure robust performance, we iterated through the train-validation-test splitting procedure fifty times. To increase predictive performance, we could use more complex methods, such as deep learning, but they require larger datasets. Studies that used deep learning or ensemble methods for similar assessments are different in terms of prediction, in which they predict the type of mutation as either pathogenic or neutral [61–64]. Even though there are a number of mutations of human CaSR in the Clinvar, the functional consequences of most of them are not known. We obtained missense variants from gnomad v4 and compared our predictions with AlphaMissense [57] and LoGoFunc [56]. While Alphamissense tends to predict benign, LoGoFunc tends to predict pathogenic outcomes even for variants that are actually neutral. LoGoFunc uses multiple features categorized into gene-level, protein-level and variant-level. They include scores of other tools, amino acid types, mode of inheritance etc. They make LoGoFunc a powerful tool to predict the pathogenicity, yet it overestimates pathogenicity for CaSR.

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 its performance. 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 consequences of a substitution is co-evolution. From the MSA 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 repeated all the steps outlined in the machine learning process. Our average AUC values for training and test among 50 replications were 0.83 and 0.77, respectively and our 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 LoF, contrary to their previous prediction of causing GoF. 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 GoF and LoF mutations.

There are other limitations that may affect the interpretation and applicability of our findings. In-species variability and potential sequencing errors present significant challenges. In-species genetic diversity can lead to varying effects of similar mutations across different individuals or populations, complicating the prediction of their functional outcomes. Furthermore, sequencing errors, which may occur during data acquisition, may cause incorrect mutation identification. These limitations can lead to flawed analyses and misinterpretations of the mutation's impact on receptor functionality. These factors highlight the necessity for

rigorous validation of genetic data and careful consideration of genetic diversity in extrapolating our results to broader applications.

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, the selection of these groups is changed, so that this process can be further automated. We did not anticipate that our specific models would match any receptors 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 the NCBI genomes website (https://ftp.ncbi.nlm.nih.gov/genomes/archive/old%20ref%20seq/) in 2018. A hmmsearch of HMMER software [65] (http://hmmer.org/) was run for each proteome against the Pfam 7tm_3 profile [66]. Sequences with significant 7tm_3 hit based on hmmsearch results (above the default threshold) were compiled from proteomes. A hmmscan of HMMER software [65] (http://hmmer.org/) was run for these sequences against the Pfam-A 32.0 database [66]. Based on the results of the 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 the given canonical proteins on the UniProt website [67]. Because plants do not have GPCRs, they were eliminated from the analysis. For single isoform sequences of each proteome, a BLAST database was built [20].

Subfamily definition and subfamily specific models

Each protein sequence of each taxon was queried through BLASTP against each prepared BLAST database [20]. 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 (the longest sequence that hit the 7tm_3). Sequences were aligned using the MAFFT v7.221 E-INS-I algorithm with default parameters [68]. A maximum likelihood based phylogenetic 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 [69]. The most common highest taxonomic level was added to the phylogenetic tree with the ETE toolkit [70]. Based on the phylogenetic tree, sequences belonging to the corresponding subfamily were taken, and a profile HMM was built. The subfamily assignment process begins by scanning each sequence with a 7tm_3 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 [68] and trimmed using the gappy-out method of the trimAl tool [71].

Paralog filter

There were a number of duplications in the CaSR subfamily. For example, Dipodomys ordii has 116 CaSR sequences. To reduce the number of sequences, human CaSR and other human class C GPCR proteins sequences compiled with CaSR sequences of each taxon and aligned with the MAFFT v7.221 auto algorithm [68], and the gappy-out method of the trimAl tool was used to trim the MSA [71]. The ML tree was built using RAxML-NG v0.9.0 with ML tree search and bootstrapping (Felsenstein Bootstrap and Transfer Bootstrap) parameters [72]. 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 to the GPRC6A and TAS1Rs subfamilies, and human GPRC6A sequence was added to CaSR subfamily as an outgroup. We aligned subfamily sequences with the MAFFT v7.221 einsi algorithm [68] and built the ML trees by using RAxML-NG v0.9.0 with the JTT model parameter [72]. 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 the ETE toolkit [70]. 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 the 'ggstatsplot' package [73].

Subfamily Specific Profile HMMs

After we took all receptors from CaSR, CaSR-like, GPRC6A, and taste receptors, we aligned them by using MAFFT v7.221 auto algorithm [68]. For each subfamily we removed the positions from the MSA that correspond to a gap in the human receptor. Then, we divided the MSA into subfamily alignments. We generated an HMM profile from the gap-removed alignment of each subfamily. The "pnone" option of the HMMER package was used in the HMM profile construction step, and thus the obtained probability parameters are simply calculated by employing observed frequencies. The constructed HMM profiles for any target subfamily, even with the "pnone" option, still hit the sequences from other subfamilies. On the other hand, we aim to obtain profiles such that they can correctly identify the elements of the target subfamily while not hitting any proteins from the remaining ones. To achieve this, we use a position-weighting approach. After determining the weight per position, we update the emission probabilities by taking these weights into account. The weights are computed by a detailed analysis of the conservation dynamics of each subfamily. As the initial step, we defined the target, the closest and the rest groups based on the ML tree. By considering the distance between the root node of each subfamily and the target one, we determined the closest group. The remaining parts of the tree were taken as the rest. To obtain specific HMM profiles for each subfamily, we considered five different scenarios that show variety in terms of the subfamilies in the target, close and rest groups.

- 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.

Our main idea is finding the positions that can help discriminate the target family from the others and assigning a high weight to these positions. Similarly, we assign a low weight to the positions that are conserved for all subfamilies since they can cause incorrect hits. The details of how we compute the weight per position is given in Algorithms 1, Algorithm 2 and S2 Fig.

In Step 1 of Algorithm 1, we show how to select the representative amino acids per group and assign a score to each group based on the amino acid frequencies obtained from MSA for any 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, close and 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 the amino acid is conserved and the value decreases with respect to the closeness of substituted amino acids.

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

After determining the types and initial score of each position, the next and final step is to decide the final weight that will be used to modify the emission probabilities of the default HMM profiles. Details of this process are given in Algorithm 2. The weight of each type will be in the following order from highest to lowest: Type I, Type IV, Type III and Type II.

The maximum or minimum weight for each category is predefined through an empirical process over the sequences that are used in our algorithm design process. As mentioned in the Results section, the sequences tested in our approach were from different species and they were not used in calculating the position weights. Type I refers to positions that show the most subfamily-specific patterns, which is why we assign a weight greater than or equal to $c_1 (c_1 \geq 1)$ to the sequences in this category. The position with the smallest initial score takes c_1, and other positions take a weight proportional to their ratio to the smallest initial score in Type I.

Specifically, $c_1 = 1$ for CaSR, GPRC6A, TAS1R1 and TAS1R3 and $c_1 = 1.5$ for TAS1R2.

Type IV positions are expected to contribute the most after Type I. We determine a maximum weight value, c_2, for this category. The maximum weight of Type IV positions, c_2, is equal to the mean value of the weights of Type I positions (line 4 of Algorithm 2) for all target subfamilies. Positions labeled as Type IV show patterns that can be used to discriminate subfamilies. However, since the conservation level at these positions can be low, we want to assign a high weight that does not exceed that of Type I. The highest initial score in this category can be as informative as Type I, so we assign the average value of Type I to that position. Other positions in this category take a weight depending on their ratio to the position with the highest score in Type IV.

The maximum weight of Type III positions, c_3, is taken as the minimum weight of Type IV positions for GPRC6A and 0.5 for TAS1Rs. For CaSR, since it is a highly conserved family and

Type I and IV positions are much more important compared to the other types, we decreased the maximum weight of Type III positions further by taking the minimum of Type IV over 2.

Finally, Type II positions take the lowest score as the highest weight, c_4, is again determined by considering the conservation pattern of the target. We took $c_4 = 0.2$ for GPRC6A and TAS1R2, $c_4 = 0.25$ for TAS1R1 and TAS1R3. For CaSR, the highest weight of Type II positions is restricted by the minimum weight of Type III positions over 2. As mentioned earlier, this type includes positions that show similar conservation patterns and cannot be used to distinguish between the elements of the target and other groups. However, they are necessary to detect sequences belonging to any of these five subfamilies, so we assign them a minimal contribution to the score. Other positions in the Type III and Type II categories take a weight depending on their ratio to the position with the highest initial score in Type III and II, respectively, as shown in Algorithm 2.

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∈{c,r}
if n_j = 1
$R_j = a_k$ where k is the subfamily in group j
$S_j = F_k$
else
if $R_T \in \{a_j, j = 1,...,n_j\}$
$R_j = R_T$
$S_j = F_k$ 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
$score_k = \Sigma_{i\in\{T,c,r\}} S_i$

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

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
$score_k = \Sigma_{i \in \{T,c.r\}} S_i$
else if $Blosum(R_T, R_C) \leq thr_{bls}$ and $Blosum(R_T, R_r) \leq thr_{bls}$
$type_k$ = I
$score_k = \Sigma_{i \in \{T,c,r\}} S_i$
else
$type_k$ = IV
$score_k = \Sigma_{i \in \{T,c,r\}} S_i$
else if R_c is gap
if $S_T \geq thr_1$ and $S_r \geq thr_2$
$type_k$ = I
$score_k = \Sigma_{i \in \{T,c,r\}} S_i$
else if $Blosum(R_T, R_r) \leq thr_{bls}$
$type_k$ = I
$score_k = \Sigma_{i \in \{T,c,r\}} S_i$
else
$type_k$ = IV
$score_k = \Sigma_{i \in \{T,c,r\}} S_i$
else if R_r is gap
if $S_T \geq thr_1$ and Sc $\geq thr_2$
$type_k$ = I
$score_k = \Sigma_{i \in \{T,c,r\}} S_i$
else if $Blosum(R_T, R_c) \leq thr_{bls}$
$type_k$ = I
$score_k = \Sigma_{i \in \{T,c,r\}} S_i$
else
$type_k$ = IV
$score_k = \Sigma_{i \in \{T,c,r\}} S_i$

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
$score_k = \Sigma_{i \in \{T,c,r\}} S_i$
else
if $Blosum(R_T, R_c) \leq thr_{bls}$ and $S_c \geq thr_2$
$type_k$ = III
$score_k = \Sigma_{i \in \{T,c,r\}} S_i$
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
$score_k = \Sigma_{i \in \{T,c,r\}} S_i$
else if $Blosum(R_T, R_c) \leq thr_{bls}$
$type_k$ = I
$score_k = \Sigma_{i \in \{T,c,r\}} S_i$

78 *else*
79 $type_k$ = *II*
80 $score_k = \frac{1}{\sum_{i\in\{T,\varepsilon,r\}} S_i}$

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

1 *for* p_1 = 1:n_1
2 $weight_{p_1} = \frac{score_{1,p_1}}{\min_{l=1,\ldots,n_1}(score_{1,l})} c_1$

Weight of Type IV positions

3 *for* p_2 = 1:n_2
4 $c_2 = mean(weight_{p_1})_{p_1\in\{1,\ldots,n_1\}}$
$weight_{p_2} = \frac{score_{2,p_2}}{\max_{l=1,\ldots,n_2}(score_{2,l})} c_2$

Weight of Type III positions

5 *for* p_3 = 1:n_3
6 $weight_{p_3} = \frac{score_{3,p_3}}{\max_{l=1,\ldots,n_3}(score_{3,l})} c_3$

Weight of Type II positions

7 *for* p_4 = 1:n_4
8 $weight_{p_4} = \frac{score_{4,p_4}}{\max_{l=1,\ldots,n_4}(score_{4,l})} c_4$
9 *if S = = GPRC6A*
$weight_{p1}$ = $weight_{p1}$*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 [74] with automatic model selection [75] (-m MFP) and ultrafast bootstrap [76] (-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 [74]. 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 [77] for GPRC6A and taste receptors since they do not have experimental structures. To visualize structures and residues, we used the UCSF Chimera tool [78].

We calculated SDP scores by a method extended from [29] by considering phylogenetic trees and a phylogeny-based scoring approach, adjPHACT, based on the methodology of the PHACT algorithm. The details of how we compute the SDP score for any position k can be found in Algorithm 3. PHACT examines the evolutionary history of proteins from the maximum likelihood phylogenetic tree and predicts the pathogenicity of a substitution. It uses ML tree and probabilities of observing amino acids at the tree nodes from the ancestral sequence reconstruction to determine the number of substitutions occurred through evolution. Additionally, it uses the branch lengths to assess the evolutionary closeness of species. PHACTcomputes the tolerance for each amino acid for the query species (human) through a tree traversal approach. By checking the probability differences, PHACT detects the location of amino acid substitutions and computes the weighted sum of positive probability differences based on the

distance between the node of change and human. On the 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 the 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 and the representative amino acid of the corresponding subfamily. Then, similar to [29] we check whether the same amino acid is conserved across all subfamilies. On the other hand, our approach differs from [29] 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 2–7). 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 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, $P^T_{a_T}$; adjPHACT score of a_T in subfamily I (i = 1,...,n), $P^i_{a_T}$; 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, $P^i_{a_i}$.

1 *for i = 1:n*

2 *Check whether the subfamily i is conserved*

$cons = -(P^i_{a_i} \leq 0.5) + (P^i_{a_i} > 0.5)$

3 *Specificity contribution of each subfamily* $S_i = 0$

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

5 $S_i = -\exp(P^i_{a_T})$

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

7 $S_i = \exp(1) - \exp(P^i_{a_i})$

8 *end*

9 *The overall weight,*

$\omega = 1 - \max_{i=1,...,n}(P^i_{a_T})$.

10 *The SDP score for position k,*

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

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 [68]. We built the ML tree by RAxML-NG-0.9.0 with the model JTT and transfer bootstrap expectation–bs-metric fbp, tbe parameters [72]. We merged the ML trees of CaSR, GPRC6A and taste receptors by checking clades using the ETE toolkit [70].

Identification of the CaSR Activation Network

To reveal the network that is important for CaSR activation, we measured changes in contact scores between residues between active and inactive receptor states. For a single protomer, we manipulated the PDB files to represent each atom as a single residue and the modified the RRCS algorithm [58] to process the manipulated PDB files, to be able to detect changes in atom level for identical residues (python codes are provided). We applied a t-test to identify significant changes and chose a p-value threshold of 0.01. Then we later combined atomic-level changes observed within a protomer for each residue to build the activation network. If atoms of a residue are involved in multiple significant changes, this is represented as multiple edges in the network, even if the residue pair is the same.

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

In both of our analyses, we used same set of 7 active state (PDB IDs 7SIL[79], 7SIM[79], 7E6T[3], 7M3G[8], 7M3F[8], 7DTT[5], 7DTV[5]) and 5 inactive state structures (7SIN[9], 7E6U[3], 7M3E[8], 7M3J[8], 7DTW[5]) human CaSR structures.

Although we calculated adjusted p-values for both analyses, we chose not to use them in this study. This decision was made because filtering based on adjusted p-values significantly reduced the number of residues we could cover. Given the exploratory nature of our research, we decided to proceed with regular p-values.

Machine learning

XGBoost is a scalable end-to-end tree boosting system where it adds a new tree to correct the prediction errors made by previous trees. It uses a gradient descent algorithm to minimize the loss when adding new models and it can be used for both regression and classification problems.

Dataset and feature preparation

To predict the consequence of a substitution in human CaSR, we used a gradient boosting-based machine learning algorithm, XGBoost [35]. We used the XGBoost library for R [80] to train our model. We selected a total of 337 LoF and GoF 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 the LoF to GoF ratio almost the same in the datasets [81]. 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 by BLOSUM62-encoded matrices. Amino acid physico-chemical feature values, including Zimmerman polarity for distribution of polar and nonpolar residues [82], average flexibility indicating structural dynamics tendency [83], Dayhoff for substitution frequency in evolution [84], average buried area as a protein compactness indicator [85], Doolittle hydropathicity for hydrophobicity characterization [86], atomic weight ratio for elemental composition [87], molecular weight for total mass, and bulkiness indicating spatial size from the ProtScale database [88]; and domain information of the reference amino acid were used as other features.

We normalized the physico-chemical feature values such as hydrophobicity, charge, and size prior to model training. This normalization involved scaling feature values to have a mean of zero and a standard deviation of one to ensure that all features contribute equally to the model, preventing features with larger numerical ranges from disproportionately influencing the model. We calculated the mean and standard deviation for each feature in the training data and used these statistics to scale the corresponding feature values in both the training and test datasets. For the training data, we calculated the mean and standard deviation for each feature. For example, Zimmermann polarity values range from 51.600 for histidine to 0.30 for leucine. We normalized them using the mean and standard deviation of the training data. Similarly, other features such as average flexibility, Dayhoff substitution frequencies, average buried area, Doolittle hydropathicity, atomic weight ratio, molecular weight, and bulkiness were scaled to match the same criteria. For the test data, we applied the mean and standard deviation derived from the training data to ensure consistency and prevent data leakage. This process ensured that all feature values were on a comparable scale, improving the robustness and accuracy of the model. 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 to 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 the nrounds parameter to 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 give an estimate of how much feature contributed to the prediction of the model made [89]. We calculated SHAP values for our final model trained by all samples by the using R shapviz package [90].

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 the 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 GoF or LoF class for any substitution. We did a literature search to find

clinical cases that cause either GoF or LoF mutations and not seen in our model. We reported our predictions in the Table 2.

Supporting information

S1 Fig. Heatmap Comparing Classification Metrics of Different Mutation Prediction Models. The performance of three prediction models—Our Model, *LoGoFunc*, and AlphaMissense—across four key metrics: Accuracy, Precision, Recall, and F1 Score are shown. The values are color-coded, with darker shades indicating higher performance. Our Model demonstrates superior performance in all metrics compared to the other tools, highlighting its enhanced predictive capability for classifying CASR gene mutations.
(TIF)

S2 Fig. Subfamily Specific Profile HMM Emission Weights. We considered different types to weight emission probabilities of profile HMMs.
(PDF)

S1 Table. Comparison of accuracy and F1 score between tools.
(PDF)

S2 Table. Predictions for all possible substitutions of Calcium-Sensing Receptor.
(CSV)

S3 Table. Predictions for the 922 reported missense mutations in GnomAD by our tool, Alphamissense and LoGoFunc.
(CSV)

Author Contributions

Conceptualization: Ogün Adebali.

Data curation: Aylin Bircan.

Formal analysis: Aylin Bircan, Nurdan Kuru, Onur Dereli, Berkay Selçuk.

Funding acquisition: Ogün Adebali.

Investigation: Aylin Bircan.

Methodology: Aylin Bircan, Nurdan Kuru.

Project administration: Ogün Adebali.

Software: Aylin Bircan, Nurdan Kuru.

Supervision: Ogün Adebali.

Visualization: Aylin Bircan.

Writing – original draft: Aylin Bircan.

Writing – review & editing: Aylin Bircan, Nurdan Kuru, Onur Dereli, Berkay Selçuk, Ogün Adebali.

References

1. Cook AE, Mistry SN, Gregory KJ, Furness SGB, Sexton PM, Scammells PJ, et al. Biased allosteric modulation at the CaS receptor engendered by structurally diverse calcimimetics. British journal of pharmacology. 2015; 172(1):185–200. https://doi.org/10.1111/bph.12937 PMID: 25220431

2. Gorvin CM. Molecular and clinical insights from studies of calcium-sensing receptor mutations. J Mol Endocrinol. 2019; 63(2):R1–R16. https://doi.org/10.1530/JME-19-0104 PMID: 31189130
3. Chen X, Wang L, Cui Q, Ding Z, Han L, Kou Y, et al. Structural insights into the activation of human calcium-sensing receptor. Elife. 2021; 10. Epub 2021/09/02. https://doi.org/10.7554/eLife.68578 PMID: 34467854; PubMed Central PMCID: PMC8476121.
4. Geng Y, Mosyak L, Kurinov I, Zuo H, Sturchler E, Cheng TC, et al. Structural mechanism of ligand activation in human calcium-sensing receptor. Elife. 2016; 5. Epub 2016/07/20. https://doi.org/10.7554/eLife.13662 PMID: 27434672; PubMed Central PMCID: PMC4977154.
5. Ling S, Shi P, Liu S, Meng X, Zhou Y, Sun W, 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):383–94. Epub 2021/02/20. https://doi.org/10.1038/s41422-021-00474-0 PMID: 33603117; PubMed Central PMCID: PMC8115157.
6. Wootten D, Christopoulos A, Marti-Solano M, Babu MM, Sexton PM. Mechanisms of signalling and biased agonism in G protein-coupled receptors. Nat Rev Mol Cell Biol. 2018; 19(10):638–53. Epub 2018/08/15. https://doi.org/10.1038/s41580-018-0049-3 PMID: 30104700.
7. Zhang C, Zhang T, Zou J, Miller CL, Gorkhali R, Yang JY, 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):e1600241. Epub 2016/07/08. https://doi.org/10.1126/sciadv.1600241 PMID: 27386547; PubMed Central PMCID: PMC4928972.
8. Gao Y, Robertson MJ, Rahman SN, Seven AB, Zhang C, Meyerowitz JG, et al. Asymmetric activation of the calcium-sensing receptor homodimer. Nature. 2021; 595(7867):455–9. Epub 2021/07/02. https://doi.org/10.1038/s41586-021-03691-0 PMID: 34194040.
9. Park J, Zuo H, Frangaj A, Fu Z, Yen LY, Zhang Z, et al. Symmetric activation and modulation of the human calcium-sensing receptor. Proc Natl Acad Sci U S A. 2021; 118(51). Epub 2021/12/18. https://doi.org/10.1073/pnas.2115849118 PMID: 34916296; PubMed Central PMCID: PMC8713963.
10. Wen T, Wang Z, Chen X, Ren Y, Lu X, Xing Y, et al. Structural basis for activation and allosteric modulation of full-length calcium-sensing receptor. Science Advances. 2021; 7(23):eabg1483. https://doi.org/10.1126/sciadv.abg1483 PMID: 34088669
11. Mun H-C, Culverston EL, Franks AH, Collyer CA, Clifton-Bligh RJ, Conigrave AD. 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):29067–72. https://doi.org/10.1074/jbc.M500002200 PMID: 15888439
12. Zhang C, Huang Y, Jiang Y, Mulpuri N, Wei L, Hamelberg D, 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):5296–309. https://doi.org/10.1074/jbc.M113.537357 PMID: 24394414
13. Zhang Z, Qiu W, Quinn SJ, Conigrave AD, Brown EM, Bai M. 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):33727–35. https://doi.org/10.1074/jbc.M200976200 PMID: 12095982
14. Liu H, Yi P, Zhao W, Wu Y, Acher F, Pin J-P, et al. Illuminating the allosteric modulation of the calcium-sensing receptor. Proceedings of the National Academy of Sciences. 2020; 117(35):21711–22. https://doi.org/10.1073/pnas.1922231117 PMID: 32817431
15. Flock T, Hauser AS, Lund N, Gloriam DE, Balaji S, Babu MM. Selectivity determinants of GPCR-G-protein binding. Nature. 2017; 545(7654):317–22. Epub 2017/05/11. https://doi.org/10.1038/nature22070 PMID: 28489817; PubMed Central PMCID: PMC5846738.
16. Chagoyen M, Garcia-Martin JA, Pazos F. Practical analysis of specificity-determining residues in protein families. Brief Bioinform. 2016; 17(2):255–61. Epub 2015/07/05. https://doi.org/10.1093/bib/bbv045 PMID: 26141829.
17. Studer RA, Dessailly BH, Orengo CA. Residue mutations and their impact on protein structure and function: detecting beneficial and pathogenic changes. Biochemical journal. 2013; 449(3):581–94. https://doi.org/10.1042/BJ20121221 PMID: 23301657
18. Koonin EV. Orthologs, paralogs, and evolutionary genomics. Annu Rev Genet. 2005; 39:309–38. Epub 2005/11/16. https://doi.org/10.1146/annurev.genet.39.073003.114725 PMID: 16285863.
19. Kuru N, Dereli O, Akkoyun E, Bircan A, Tastan O, Adebali O. PHACT: Phylogeny-Aware Computing of Tolerance for Missense Mutations. Mol Biol Evol. 2022; 39(6). Epub 2022/06/01. https://doi.org/10.1093/molbev/msac114 PMID: 35639618; PubMed Central PMCID: PMC9178230.
20. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. Journal of molecular biology. 1990; 215(3):403–10. https://doi.org/10.1016/S0022-2836(05)80360-2 PMID: 2231712

21. Pin JP, Galvez T, Prezeau L. Evolution, structure, and activation mechanism of family 3/C G-protein-coupled receptors. Pharmacol Ther. 2003; 98(3):325–54. Epub 2003/06/05. https://doi.org/10.1016/s0163-7258(03)00038-x PMID: 12782243.

22. Harpsoe K, Boesgaard MW, Munk C, Brauner-Osborne H, Gloriam DE. Structural insight to mutation effects uncover a common allosteric site in class C GPCRs. Bioinformatics. 2017; 33(8):1116–20. Epub 2016/12/25. https://doi.org/10.1093/bioinformatics/btw784 PMID: 28011766; PubMed Central PMCID: PMC5408886.

23. Goes van Naters W, Mucignat-Caretta C. Frontiers in Neuroscience Drosophila Pheromones: From Reception to Perception. Neurobiology of Chemical Communication. Boca Raton (FL): CRC Press/Taylor & Francis (c). 2014.

24. Brown D, Krishnamurthy N, Dale JM, Christopher W, Sjolander K. Subfamily hmms in functional genomics. Pac Symp Biocomput. 2005:322–33. Epub 2005/03/12. PMID: 15759638.

25. Srivastava PK, Desai DK, Nandi S, Lynn AM. 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; 8:104. Epub 2007/03/29. https://doi.org/10.1186/1471-2105-8-104 PMID: 17389042; PubMed Central PMCID: PMC1852395.

26. Roberts MS, Gafni RI, Brillante B, Guthrie LC, Streit J, Gash D, et al. Treatment of Autosomal Dominant Hypocalcemia Type 1 With the Calcilytic NPSP795 (SHP635). J Bone Miner Res. 2019; 34(9):1609–18. Epub 2019/05/08. https://doi.org/10.1002/jbmr.3747 PMID: 31063613; PubMed Central PMCID: PMC6744344.

27. Pruitt KD, Tatusova T, Maglott DR. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. 2007; 35(Database issue):D61–5. Epub 2006/11/30. https://doi.org/10.1093/nar/gkl842 PMID: 17130148; PubMed Central PMCID: PMC1716718.

28. Gorvin CM. Calcium-sensing receptor signaling—How human disease informs biology. Current opinion in endocrine and metabolic research. 2021; 16:10–8. https://doi.org/10.1016/j.coemr.2020.06.007 PMID: 34141952

29. Bradley D, Beltrao P. Evolution of protein kinase substrate recognition at the active site. PLoS biology. 2019; 17(6):e3000341. https://doi.org/10.1371/journal.pbio.3000341 PMID: 31233486

30. Chun L, Zhang WH, Liu JF. Structure and ligand recognition of class C GPCRs. Acta Pharmacol Sin. 2012; 33(3):312–23. Epub 2012/01/31. https://doi.org/10.1038/aps.2011.186 PMID: 22286915; PubMed Central PMCID: PMC4077135.

31. Pi M, Nishimoto SK, Quarles LD. GPRC6A: Jack of all metabolism (or master of none). Mol Metab. 2017; 6(2):185–93. Epub 2017/02/10. https://doi.org/10.1016/j.molmet.2016.12.006 PMID: 28180060; PubMed Central PMCID: PMC5279936.

32. Nango E, Akiyama S, Maki-Yonekura S, Ashikawa Y, Kusakabe Y, Krayukhina E, et al. Taste substance binding elicits conformational change of taste receptor T1r heterodimer extracellular domains. Sci Rep. 2016; 6:25745. Epub 2016/05/11. https://doi.org/10.1038/srep25745 PMID: 27160511; PubMed Central PMCID: PMC4861910 taste-substance evaluation method by use of the T1rLBD FRET assay conditions.

33. Kumari R, Castillo C, Francesconi A. Agonist-dependent signaling by group I metabotropic glutamate receptors is regulated by association with lipid domains. J Biol Chem. 2013; 288(44):32004–19. Epub 2013/09/21. https://doi.org/10.1074/jbc.M113.475863 PMID: 24045944; PubMed Central PMCID: PMC3814796.

34. Nuemket N, Yasui N, Kusakabe Y, Nomura Y, Atsumi N, Akiyama S, et al. Structural basis for perception of diverse chemical substances by T1r taste receptors. Nat Commun. 2017; 8:15530. Epub 2017/05/24. https://doi.org/10.1038/ncomms15530 PMID: 28534491; PubMed Central PMCID: PMC5457512.

35. Chen T, Guestrin C, editors. Xgboost: A scalable tree boosting system. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining; 2016.

36. Chou KJ, Hsu CY, Huang CW, Chen HJ, Ou SH, Chen CL, 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):113080. Epub 2022/02/23. https://doi.org/10.1016/j.yexcr.2022.113080 PMID: 35192837.

37. Moon JE, Yang HY, Wee G, Par KS, Ko CW. A cell function study on calcium regulation of a novel calcium-sensing receptor mutation (p.Tyr825Phe). Ann Pediatr Endocrinol Metab. 2021; 26(1):24–30. Epub 2020/09/03. https://doi.org/10.6065/apem.2040022.011 PMID: 32871647; PubMed Central PMCID: PMC8026336.

38. Palmieri S, Grassi G, Guarnieri V, Chiodini I, Arosio M, Eller-Vainicher C. Case Report: Unusual Presentations of Loss-of-Function Mutations of the Calcium-Sensing Receptor. Front Med (Lausanne).

2021; 8:809067. Epub 2022/02/11. https://doi.org/10.3389/fmed.2021.809067 PMID: 35141253; PubMed Central PMCID: PMC8818680.

39. Li N, Li X, Ni XL, Li XY, Xia WB, Yang GQ, 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):1869–71. Epub 2021/08/17. https://doi.org/10.1097/CM9.0000000000001568 PMID: 34397587; PubMed Central PMCID: PMC8367063.

40. Hao Y, Lei Z, Shi N, Yu L, Ji W, Zhang X. Radiofrequency Ablation of Parathyroid Glands to Treat a Patient With Hypercalcemia Caused by a Novel Inactivating Mutation in CaSR. Front Endocrinol (Lausanne). 2021; 12:743517. Epub 2022/02/01. https://doi.org/10.3389/fendo.2021.743517 PMID: 35095753; PubMed Central PMCID: PMC8795859.

41. Tsuji T, Hiroyuki A, Uraki S, Doi A, Morita S, Iwakura H, et al. Autosomal Dominant Hypocalcemia With Atypical Urine Findings Accompanied by Novel CaSR Gene Mutation and VitD Deficiency. J Endocr Soc. 2021; 5(3):bvaa190. Epub 2021/01/29. https://doi.org/10.1210/jendso/bvaa190 PMID: 33506158; PubMed Central PMCID: PMC7814383.

42. Dharmaraj P, Gorvin CM, Soni A, Nelhans ND, Olesen MK, Boon H, et al. Neonatal Hypocalcemic Seizures in Offspring of a Mother With Familial Hypocalciuric Hypercalcemia Type 1 (FHH1). J Clin Endocrinol Metab. 2020; 105(5). Epub 2020/03/10. https://doi.org/10.1210/clinem/dgaa111 PMID: 32150253; PubMed Central PMCID: PMC7096312.

43. Gomes V, Silvestre C, Ferreira F, Bugalho M. Autosomal dominant hypocalcaemia: identification of two novel variants of CASR gene. BMJ Case Rep. 2020; 13(6). Epub 2020/06/10. https://doi.org/10.1136/bcr-2020-234391 PMID: 32513763; PubMed Central PMCID: PMC7282295.

44. Magno AL, Leatherbarrow KM, Brown SJ, Wilson SG, Walsh JP, Ward BK. Functional Analysis of Calcium-Sensing Receptor Variants Identified in Families Provisionally Diagnosed with Familial Hypocalciuric Hypercalcaemia. Calcif Tissue Int. 2020; 107(3):230–9. Epub 2020/07/09. https://doi.org/10.1007/s00223-020-00715-1 PMID: 32638038.

45. Majumdar SK, Jacob T, Bale A, Bailey A, Kwon J, Hughes T, et al. A Novel Variant in the Calcium-Sensing Receptor Associated with Familial Hypocalciuric Hypercalcemia and Low-to-Normal PTH. Case Rep Endocrinol. 2020; 2020:8752610. Epub 2020/10/17. https://doi.org/10.1155/2020/8752610 PMID: 33062349; PubMed Central PMCID: PMC7555459 the reporting of the case or publication of this article.

46. Zajickova K, Dvorakova M, Moravcova J, Vcelak J, Goltzman D. 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):S321–S8. Epub 2020/10/24. https://doi.org/10.33549/physiolres.934522 PMID: 33094630; PubMed Central PMCID: PMC8603734.

47. Sagi SV, Joshi H, Trotman J, Elsey T, Swamy A, Rajkanna J, et al. A novel CASR variant in a family with familial hypocalciuric hypercalcaemia and primary hyperparathyroidism. Endocrinol Diabetes Metab Case Rep. 2020; 2020. Epub 2021/01/13. https://doi.org/10.1530/EDM-20-0084 PMID: 33434173; PubMed Central PMCID: PMC7576638.

48. Dong Q, Song F, Du M, Qiu M, Chen X. [Clinical and genetic analysis of a child with neonatal severe parathyroidism]. Zhonghua Yi Xue Yi Chuan Xue Za Zhi. 2020; 37(11):1247–9. Epub 2020/11/13. https://doi.org/10.3760/cma.j.cn511374-20191118-00587 PMID: 33179231.

49. Wejaphikul K, Dejkhamron P, Khorana J, Watcharachan K, Intachai W, Olsen B, et al. Subtotal parathyroidectomy successfully controls calcium levels of patients with neonatal severe hyperparathyroidism carrying a novel CASR mutation. Hormone Research in Paediatrics. 2023:1–7. https://doi.org/10.1159/000528568 PMID: 36626889

50. Courtney A, Hill A, Smith D, Agha A. 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):324. https://doi.org/10.1186/s12902-022-01231-z PMID: 36536367

51. Bletsis P, Metzger R, Nelson JA, Gasparini J, Alsayed M, Milas M. A Novel missense CASR gene sequence variation resulting in familial hypocalciuric hypercalcemia. AACE Clinical Case Reports. 2022; 8(5):194–8. https://doi.org/10.1016/j.aace.2022.05.002 PMID: 36189134

52. Wu Y, Zhang C, Huang X, Cao L, Liu S, Zhong P. Autosomal dominant hypocalcemia with a novel CASR mutation: a case study and literature review. Journal of International Medical Research. 2022; 50(7):03000605221110489. https://doi.org/10.1177/03000605221110489 PMID: 35818129

53. Coughlan A, Khan F, Brassill M. A Novel Genetic Variant Resulting in Familial Hypocalciuric Hypercalcaemia. Irish Medical Journal. 2022; 115(2):545–. PMID: 35420006

54. Mullin BH, Pavlos NJ, Brown SJ, Walsh JP, McKellar RA, Wilson SG, et al. Functional assessment of calcium-sensing receptor variants confirms familial hypocalciuric hypercalcemia. Journal of the Endocrine Society. 2022; 6(5):bvac025. https://doi.org/10.1210/jendso/bvac025 PMID: 35356007

55. Chen S, Francioli LC, Goodrich JK, Collins RL, Kanai M, Wang Q, et al. A genomic mutational constraint map using variation in 76,156 human genomes. Nature. 2024; 625(7993):92–100. https://doi.org/10.1038/s41586-023-06045-0 PMID: 38057664

56. Stein D, Kars ME, Wu Y, Bayrak Ç S, Stenson PD, Cooper DN, et al. Genome-wide prediction of pathogenic gain- and loss-of-function variants from ensemble learning of a diverse feature set. Genome Med. 2023; 15(1):103. Epub 2023/12/01. https://doi.org/10.1186/s13073-023-01261-9 PMID: 38037155; PubMed Central PMCID: PMC10688473.

57. Cheng J, Novati G, Pan J, Bycroft C, Zemgulyte A, Applebaum T, et al. Accurate proteome-wide missense variant effect prediction with AlphaMissense. Science. 2023; 381(6664):eadg7492. Epub 2023/09/21. https://doi.org/10.1126/science.adg7492 PMID: 37733863.

58. Zhou Q, Yang D, Wu M, Guo Y, Guo W, Zhong L, et al. Common activation mechanism of class A GPCRs. Elife. 2019; 8. Epub 2019/12/20. https://doi.org/10.7554/eLife.50279 PMID: 31855179; PubMed Central PMCID: PMC6954041.

59. Hauser AS, Kooistra AJ, Munk C, Heydenreich FM, Veprintsev DB, Bouvier M, et al. GPCR activation mechanisms across classes and macro/microscales. Nat Struct Mol Biol. 2021; 28(11):879–88. Epub 2021/11/12. https://doi.org/10.1038/s41594-021-00674-7 PMID: 34759375; PubMed Central PMCID: PMC8580822 a founder and a director of Z7 Biotech Ltd. After the completion of this study, C.M. moved to become an employee of Novozymes A/S. The other authors declare no competing interests.

60. Jones EM, Lubock NB, Venkatakrishnan A, Wang J, Tseng AM, Paggi JM, et al. Structural and functional characterization of G protein–coupled receptors with deep mutational scanning. Elife. 2020; 9: e54895. https://doi.org/10.7554/eLife.54895 PMID: 33084570

61. Alirezaie N, Kernohan KD, Hartley T, Majewski J, Hocking TD. ClinPred: prediction tool to identify disease-relevant nonsynonymous single-nucleotide variants. The American Journal of Human Genetics. 2018; 103(4):474–83. https://doi.org/10.1016/j.ajhg.2018.08.005 PMID: 30220433

62. Ioannidis NM, Rothstein JH, Pejaver V, Middha S, McDonnell SK, Baheti S, et al. REVEL: an ensemble method for predicting the pathogenicity of rare missense variants. The American Journal of Human Genetics. 2016; 99(4):877–85. https://doi.org/10.1016/j.ajhg.2016.08.016 PMID: 27666373

63. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic acids research. 2019; 47(D1):D886–D94. https://doi.org/10.1093/nar/gky1016 PMID: 30371827

64. Rogers MF, Shihab HA, Mort M, Cooper DN, Gaunt TR, Campbell C. FATHMM-XF: accurate prediction of pathogenic point mutations via extended features. Bioinformatics. 2018; 34(3):511–3. https://doi.org/10.1093/bioinformatics/btx536 PMID: 28968714

65. Eddy SR. Accelerated profile HMM searches. PLoS computational biology. 2011; 7(10):e1002195. https://doi.org/10.1371/journal.pcbi.1002195 PMID: 22039361

66. Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, et al. The Pfam protein families database: towards a more sustainable future. Nucleic acids research. 2016; 44(D1):D279–D85. https://doi.org/10.1093/nar/gkv1344 PMID: 26673716

67. Consortium TU. UniProt: the universal protein knowledgebase. Nucleic acids research. 2017; 45(D1): D158–D69. https://doi.org/10.1093/nar/gkw1099 PMID: 27899622

68. Yamada KD, Tomii K, Katoh K. Application of the MAFFT sequence alignment program to large data—reexamination of the usefulness of chained guide trees. Bioinformatics. 2016; 32(21):3246–51. https://doi.org/10.1093/bioinformatics/btw412 PMID: 27378296

69. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014; 30(9):1312–3. https://doi.org/10.1093/bioinformatics/btu033 PMID: 24451623

70. Huerta-Cepas J, Serra F, Bork P. ETE 3: reconstruction, analysis, and visualization of phylogenomic data. Molecular biology and evolution. 2016; 33(6):1635–8. https://doi.org/10.1093/molbev/msw046 PMID: 26921390

71. Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 2009; 25(15):1972–3. Epub 2009/06/10. https://doi.org/10.1093/bioinformatics/btp348 PMID: 19505945; PubMed Central PMCID: PMC2712344.

72. Kozlov AM, Darriba D, Flouri T, Morel B, Stamatakis A. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. Bioinformatics. 2019; 35(21):4453–5. https://doi.org/10.1093/bioinformatics/btz305 PMID: 31070718

73. Patil I. Visualizations with statistical details: The'ggstatsplot'approach. Journal of Open Source Software. 2021; 6(61):3167.

74. Nguyen L-T, Schmidt HA, Von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Molecular biology and evolution. 2015; 32(1):268–74. https://doi.org/10.1093/molbev/msu300 PMID: 25371430

75. Kalyaanamoorthy S, Minh BQ, Wong TK, Von Haeseler A, Jermiin LS. ModelFinder: fast model selection for accurate phylogenetic estimates. Nature methods. 2017; 14(6):587–9. https://doi.org/10.1038/nmeth.4285 PMID: 28481363

76. Hoang DT, Chernomor O, Von Haeseler A, Minh BQ, Vinh LS. UFBoot2: improving the ultrafast bootstrap approximation. Molecular biology and evolution. 2018; 35(2):518–22. https://doi.org/10.1093/molbev/msx281 PMID: 29077904

77. Waterhouse A, Bertoni M, Bienert S, Studer G, Tauriello G, Gumienny R, et al. SWISS-MODEL: homology modelling of protein structures and complexes. Nucleic acids research. 2018; 46(W1):W296–W303. https://doi.org/10.1093/nar/gky427 PMID: 29788355

78. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, et al. UCSF Chimera—a visualization system for exploratory research and analysis. J Comput Chem. 2004; 25(13):1605–12. Epub 2004/07/21. https://doi.org/10.1002/jcc.20084 PMID: 15264254.

79. Park J, Zuo H, Frangaj A, Fu Z, Yen LY, Zhang Z, et al. Symmetric activation and modulation of the human calcium-sensing receptor. Proceedings of the National Academy of Sciences. 2021; 118(51): e2115849118. https://doi.org/10.1073/pnas.2115849118 PMID: 34916296

80. Chen T, He T, Benesty M, Khotilovich V, Tang Y, Cho H, et al. Xgboost: extreme gradient boosting. R package version 04–2. OS Independent. 2015.

81. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. the Journal of machine Learning research. 2011; 12:2825–30.

82. Zimmerman J, Eliezer N, Simha R. The characterization of amino acid sequences in proteins by statistical methods. Journal of theoretical biology. 1968; 21(2):170–201. https://doi.org/10.1016/0022-5193(68)90069-6 PMID: 5700434

83. Bhaskaran R, Ponnuswamy P. Positional flexibilities of amino acid residues in globular proteins. International Journal of Peptide and Protein Research. 1988; 32(4):241–55.

84. Barker WC, Ketcham LK, Dayhoff MO. A comprehensive examination of protein sequences for evidence of internal gene duplication. Journal of Molecular Evolution. 1978; 10:265–81. https://doi.org/10.1007/BF01734217 PMID: 633380

85. Rose GD, Geselowitz AR, Lesser GJ, Lee RH, Zehfus MH. Hydrophobicity of amino acid residues in globular proteins. Science. 1985; 229(4716):834–8. https://doi.org/10.1126/science.4023714 PMID: 4023714

86. Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. Journal of molecular biology. 1982; 157(1):105–32. https://doi.org/10.1016/0022-2836(82)90515-0 PMID: 7108955

87. Grantham R. Amino acid difference formula to help explain protein evolution. science. 1974; 185 (4154):862–4. https://doi.org/10.1126/science.185.4154.862 PMID: 4843792

88. Walker JM. The proteomics protocols handbook: Springer; 2005.

89. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. Advances in neural information processing systems. 2017; 30.

90. Mayer M. shapviz: SHAP Visualizations. R package version 0.9.0 2023. Available from: https://github.com/ModelOriented/shapviz.