Computational pipeline to probe NaV1.7 gain-of-function variants in neuropathic painful syndromes

Alberto A. Toffano1, Giacomo Chiarot2, Stefano Zamuner3, Margherita Marchi4, Erika Salvi4, Stephen G. Waxman5, Catharina G. Faber6,7, Giuseppe Lauria4,8, Achille Giacometti1,9 & Marta Simeoni2,9

Applications of machine learning and graph theory techniques to neuroscience have witnessed an increased interest in the last decade due to the large data availability and unprecedented technology developments. Their employment to investigate the effect of mutational changes in genes encoding for proteins modulating the membrane of excitable cells, whose biological correlates are assessed at electrophysiological level, could provide useful predictive clues. We apply this concept to the analysis of variants in sodium channel NaV1.7 subunit found in patients with chronic painful syndromes, by the implementation of a dedicated computational pipeline empowering different and complementary techniques including homology modeling, network theory, and machine learning. By testing three templates of different origin and sequence identities, we provide an optimal condition for its use. Our findings reveal the usefulness of our computational pipeline in supporting the selection of candidates for cell electrophysiology assay and with potential clinical applications.

NaV1.7 is responsible of the propagation of the pain stimuli through the peripheral nervous system. It belongs to the family of Voltage Gated Sodium Channels (VGSCs) proteins expressed both in the prokaryotic and the eukaryotic realms. The most known biophysical activity is carried out by the α-subunits that, for eukarya, is formed by the three-dimensional rearrangement of a single polypeptide chain of almost 2,000 amino acids. This chain is wrapped in four domains (DI-DIV), each composed of six transmembrane helices, with the first four S1–S4 forming the voltage-sensing domain (VSD), and the last two S5–S6, as well as the extracellular linkers (P-loop) between S5 and S6, forming the selectivity filter (SF) and the central aqueous pore domain (PD), as shown in Fig. 1. The reaction to the change of membrane potential is mediated by helix S4, thanks to the positively charged amino acids arginine (R) and lysine (K) located along the segment.

Note that the positive charges are wrapped and stabilized by amino acids present in the helices from S1 to S3. Depolarization of the membrane is believed to produce a motion toward the extracellular side of S4 segments of each domain, and this motion is transferred to the pore domain via intracellular linkers between the segments S4 and S5. The conformational change results in the opening of the channel pore. The segments S5 and S6 are scaffold of the pore channel, while the extracellular linkers are responsible for the selectivity filter to Na ions that is highly conserved and composed of aspartate (D) in DI, glutamate (E) in DII, lysine (K) in DIII, and alanine (A) in DIV forming a ring, which is the geometrically narrowest region of the ion pore, see Fig. 3. In mammals, there are nine different isoforms of α-subunits, Nav1.1–Nav1.9, and their ratio of expression changes among different type of tissues. The isoform NaV1.7 we consider, which is encoded by SCN9A, is highly expressed in nociceptors, where it has a pivotal role in subthreshold membrane depolarization. Here, single aminoacid substitutions in patients diagnosed with inherited erythromelalgia (IEM), paroxysmal extreme pain disorder (PEPD), small fiber neuropathy (SFN) and painful diabetic neuropathy (PDN) were found to induce a gain-of-function of the channel.

1Dipartimento di Scienze Molecolari e Nanosistemi, Università Ca’ Foscari Venezia, Venezia-Mestre, Italy. 2Dipartimento di Scienze Ambientali, Informatica e Statistica, Università Ca’ Foscari Venezia, Venezia-Mestre, Italy. 3Laboratory of Statistical Biophysics, Institute of Physics, School of Basic Sciences, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland. 4Neuroalgology Unit, Fondazione IRCCS Istituto Neurologico “Carlo Besta”, Milan, Italy. 5Center for Neuroscience and Regeneration Research, VA Connecticut Healthcare System and Yale Medical School, West Haven, USA. 6MHeNs school for Mental Health and Neuroscience, Maastricht University, Maastricht, The Netherlands. 7Department of Neurology, Maastricht University Medical Center, Maastricht, The Netherlands. 8Department of Biomedical and Clinical Sciences “Luigi Sacco”, University of Milan, Milan, Italy. 9European Centre for Living Technology (ECLT), Venice, Italy. *email: simeoni@unive.it
A significant challenge that is posed to the above scenario is given by the large number of possible case studies, as well as the difficulties involved in performing electrophysiology studies at large scales. On the other hand, technologies such as artificial intelligence can be applied to gain insights from multiple data sources and rationalize them to improve clinical decision-making ability and improve patient outcomes. This is particularly true in neuroscience where recent findings have shown the power of these tools both as classifiers and predictors, thus envisaging the use of computational pipelines in workflows fully integrated into clinical applications. Indeed, it was recently suggested that the available computerized predictive algorithms "are not infallible" and a more holistic approach was in order.

Building upon past work by our group, in this study we provide one such computational pipeline combining homology modeling, Residue Interaction Networks (RINs) and machine learning techniques, to determine whether there is a common structural pattern linking gain-of-function pathogenic mutations to each other. Starting from a given template, we identify a set of sequences representative of some known gain-of-function pathogenic mutations (PAT), as well as other neutral (NEUTRAL) variants, and use homology modeling to generate the corresponding three-dimensional structures. In order to make the relative comparison of these structures feasible, the information of the topology of each structure is then used to map them into their corresponding graphs using RINs, thus reducing the dimensionality of the problem and hence its complexity. This step then allows for a pairwise comparison of all the obtained networks via graph kernel techniques, and for classifying them through machine learning techniques that are able to identify common structural patterns. Note that here NEUTRAL refers to genetic variants selected either from annotated sequence substitutions not known to affect a protein’s function and possibly contribute to genetic disease or from pseudo mutations between orthologous NaV1.7 proteins in closely related mammals.

Three different templates are presented as case studies. The first one is the same template used in past work and it will be used as a benchmark for our analysis. The other two templates derive from two other homologous proteins with greater identities with the original sequence. All three templates depict the sodium channel protein in its closed state and have different sequence identities. The comparative results from these three different case studies will provide useful information on the general potentialities of the proposed workflow, the optimal condition for its use, as well as the limiting factors that need to be addressed to make it an useful tool in clinical applications.

In short, the main aim of our computational workflow is to predict whether a mutation is expected to be deleterious and be able to be distinguished from all others. We then benchmark our method by applying it to the 85 SCN9A mutations reported in Fig. 4a and see whether it is able to classify PAT mutations and discriminate them from NEUTRAL ones.

It is important to emphasize that the present work considers only gain-of-function mutations and relies upon the critical assumption that the functional impact of a point mutation is caused by changes in residue-residue interaction network only, while several additional effects can be induced by point mutation and affect the functionality of the channel. Indeed ion-channels function is determined by many factors besides the shift in gating kinetics. At least 30 mutations have been described in scientific literature for impairing the Nav1.7 function, either by haploinsufficiency effect or by hindering the selective permeability of the pore. However, these kinds of null-mutations result in a lower quantity of functional Nav1.7 on the membrane, and are associated with Congenital Insensitivity to Pain (CIP) a disease characterized by pain signaling defects.
Results and discussion

Our computational pipeline is sketched in Fig. 2 in the specific case of human protein Nav1.7 as in Ref. 6, the NCBI sequence NP_002968.1 has been chosen as the Wild Type (WT) sequence for our study. It is classified as the first isoform of the transcript variant 1, identification code NM_002977.3 for the chromosomal sequence. It is composed by four domains DI-DIV as pictured in Fig. 3 (top), located at positions DI (112–410), DII (715–978), DIII (1169–1477), DIV (1488–1784). Fig. 3 (bottom) provides its pictorial representation along the sequence, and Fig. 1 its snapshot both as side (a) and top (b) views.

We then create 85 copies of the WT sequence each one with a single amino-acid mutation, divided into pathogenic (referred to as \textbf{PAT}) and non-pathogenic (henceforth \textbf{NEUTRAL}) variants. The first 31 \textbf{PAT} mutations have been proved to be associated with gain-of-function of the protein by cell electrophysiology...
### Table: MUTATION vs. NEUTRAL

| Disease | MUTATION | NEUTRAL |
|---------|----------|---------|
| IEM     |          |         |
| SFN     |          |         |
| PPD     |          |         |
| PDE     |          |         |

(a) The considered genetic mutations and corresponding ids. Color coding is according to Figure 3.

**Figure 4.** [Left] (a) MUTATION and NEUTRAL genetic variants. MUTATION mutations are further divided by disease and highlighted with different colors according to Fig. 3. Among the NEUTRAL variants, the 21 known human variants are highlighted in bold and the 4 not causing biophysical abnormalities are also starred. [Right] Mutation I136V: the initial Isoleucine (b) in position 136 is turned into a Valine (c).

As an example, Fig. 4b,c display a blow up of the three-dimensional structure on the I136V mutation, where the isoleucine in position 136 is mutated into a valine residue. Among the gain-of-function MUTATION mutations, there are 18 causing Inherited Erythromelalgia (IEM), 6 associated with Small Fibre Neuropathy (SFN), 6 causing Paroxysmal Extreme Pain Disorder (PEPD) and 1 related with painful neuropathy in diabetic patients (PDN). The list of all the considered genetic variants is shown in Fig. 4a, where the first 31 pathogenic are those involved in the represented disease, color-coded accordingly20–58. As an example, Fig. 4b,c display a blow up of the three-dimensional structure on the I136V mutation, where the isoleucine in position 136 is mutated into a valine residue.

Three different template structures, describing the closed state of the sodium channel, have been selected. We have used Clustal Omega26 to assess the sequence identity between each template and the original WT. The NavAb is derived from the cryo-EM structure of the ortholog protein of Acrobacter butzleri (PDB code 3RVY), reporting the channel captured in a closed-pore conformation with four activated voltage sensors at 2.7 Å resolution, and modified as described in Ref.6. We shall refer to it as MOESM3 hereafter. The sequence identity to Homo Sapiens is 50.8% compared to the WT sequence.

Among the gain-of-function MUTATION mutations, there are 18 causing Inherited Erythromelalgia (IEM), 6 associated with Small Fibre Neuropathy (SFN), 6 causing Paroxysmal Extreme Pain Disorder (PEPD) and 1 related with painful neuropathy in diabetic patients (PDN). The list of all the considered genetic variants is shown in Fig. 4a, where the first 31 pathogenic are those involved in the represented disease, color-coded accordingly20–58. As an example, Fig. 4b,c display a blow up of the three-dimensional structure on the I136V mutation, where the isoleucine in position 136 is mutated into a valine residue.

Three different template structures, describing the closed state of the sodium channel, have been selected. We have used Clustal Omega26 to assess the sequence identity between each template and the original WT. The NavAb is derived from the cryo-EM structure of the ortholog protein of Acrobacter butzleri (PDB code 3RVY), reporting the channel captured in a closed-pore conformation with four activated voltage sensors at 2.7 Å resolution, and modified as described in Ref.6. We shall refer to it as MOESM3 hereafter. The sequence identity to Homo Sapiens is 50.8% compared to the WT sequence.

The NavAb is derived from the cryo-EM structure of the ortholog protein of Acrobacter butzleri (PDB code 3RVY), reporting the channel captured in a closed-pore conformation with four activated voltage sensors at 2.7 Å resolution, and modified as described in Ref.6. We shall refer to it as MOESM3 hereafter. The sequence identity to Homo Sapiens is 50.8% compared to the WT sequence.

Finally, the last selected template hNav1.7 is derived from the eukaryotic protein NaV1.7 of Homo Sapiens via cryo-EM (PDB code 6J8J) and will be referred to as 6J8J. It represents an inactivated state with all four Voltage Sensitive domains in the “up” conformation and the intracellular gate closed, meaning that the channel is closed and not available for activation. The alignments of the WT sequence and this template is 97%. The detailed alignments of all used templates are reported in the Supplementary Material.
When the primary sequence and the function of the protein are known, we use Direct Coupling Analysis (DCA), homology modeling, and energy minimization to obtain the three-dimensional structure of the mutated sequence. DCA provides an assessment on the quality of a given template and hinges upon the idea that the mutation of any amino acid is constrained by the need of conserving protein function, so that only mutations conserving the original physical properties are allowed. Then, homology modeling allows the transfer of a protein's quaternary structure from one sequence to another by homology. The final step is energy minimization to remove possible steric clashes and inconsistencies that might arise during this process. We have used an in-house implementation of DCA, Swiss-Model for the homology modelling and FG-MD for energy minimization. Once the structures have been generated we use QMEANBrane and RAMPAGE to evaluate the quality of the obtained structure in terms of their Ramachandran angles. Hydrogen bond, Van der Waals, Ionic bridges, π – π Stacking, cation and Disulfide bridges. See section “Methods” for details on the above techniques.

After the quality assessment, a set of 85 three-dimensional structures, representing the 85 genetic variants reported in Fig. 4, are available and need to be compared with each other to test for differences. This is in general a very challenging task when carried out in a full fledged three-dimensional representation, where the comparison is carried out residue by residue. A possible way around is to map each structure into its graph representation. In this way we clearly loose the three-dimensional information on the spatial arrangement of each amino acid, but we can still compare the relative topologies that is crucial to identify common patterns. This step can be done via the Residue Interaction Network (RIN) technique, where nodes represent amino acids and edges represent their non-covalent interactions (see last step at the top of Fig. 2). Here, we use RING 2.0 that is able to distinguish six different types of interactions: Hydrogen bond, Van der Waals, Ionic bridges, π – π Stacking, and Disulfide bridges. See section “Methods” for more details.

The final step of the computational pipeline is to compare and classify the obtained networks, as shown in the lower part of Fig. 2. To accomplish these tasks we employ Graph Kernels to perform pairwise RINs comparison, and then both unsupervised and supervised machine learning to identify common patterns. We used two different kernels to compare RINs, the Vertex Histogram (VH) and Weisfeiler–Lehman (WL) subtree kernels, implemented by the GraKel Python library. Both of them are based on node labels comparison, where node labels are set as the position of each node—i.e. amino acid, in the protein sequence. The essential difference between the two methods is that the VH method makes a pairwise comparison of the nodes labels only, whereas the WL method compares also the corresponding subtrees, hence providing a more complete and general contrast.

As for the unsupervised learning, we consider the Dominant Set clustering method where a set of elements is organized in clusters, in such a way that each cluster satisfies an internal homogeneity and expresses an external inhomogeneity. We also apply the Average-Linkage hierarchical clustering. Supervised learning has also been considered using the Support Vector Machine (SVM) method implemented in the Scikit learn library.

Details on all the above calculations can be found in section “Methods” and in the Supplementary Material. In short, our computational pipeline (Fig. 2) includes the following steps.

**Step 1** Reconstruction of the three-dimensional structures of the mutated sequences via DCA, homology modelling, and energy minimization. See Figs. 3 and 4;

**Step 2** Quality assessment of the obtained structures. See Figs. 5 and 6;

**Step 3** Mapping of each three-dimensional structure into a network, comparison between networks via Graph Kernels techniques, and identification of pathogenic mutations via machine learning techniques. See Fig. 7.

We stress the particular importance of obtaining high quality structures from the homology modeling, because any drawbacks at this stage could affect all downstream analyses, affecting network representations and making comparison and classification questionable. As we shall see below, however, this step is quite delicate. Indeed other possible templates, in addition to those reported here, could have been selected in principle. For instance, Huang and collaborators used the recently determined atomic structure of the rabbit voltage-gated calcium channel Cav1.1 as template for Nav1.7 studies. Moreover, other human templates have also been proposed recently. However, as detailed in the Supplementary Material (see Figures S7 and S8), it turns out that they could not be used in our case for either low resolution in the original crystal or for a limited extension compared with our WT.

While the reliability of the MOESM3 has already been assessed, those of 6A90 and 6J8J have not. As a preliminary step, we then first performed a DCA analysis of the 6A90 and 6J8J templates whose predictions are shown in Fig. 5.

In particular, Fig. 5 displays the comparison of the contact maps for the 6A90 and 6J8J templates (green points) with the corresponding DCA predictions (red points). Fig. 5a,b show this comparison for the full sequence of the 6A90 and the 6J8J, respectively. Here, the alignment contains a relatively small number of residues and quite large fraction of gaps, and the predicted contact map is relatively noisy. On the other hand, if we restrict the analysis to the region of the alignment for which the fraction of gaps is less than 30%, which mainly correspond to the region from residue 1181 to residue 1851, then the predictions display a good agreement with the deposited structures 6A90 and 6J8J as illustrated in Fig. 5c,d. Several reasons could explain the above discrepancy between the experimental deposited structure and DCA predictions. Firstly, the predicted rearrangement could be due to an allosteric conformation of the protein different from the one seen in the experiments. Secondly, these interactions could be important for a large fraction of the sequences in our alignment, but not necessarily for the crystallized protein. Finally, the discrepancy could be due to crystal artifacts, not uncommon for membrane proteins.
Bearing this information in mind, we next proceed to the quality analysis of the WT template 6A90 with QMEANBrane and RAMPAGE, as shown in Fig. 6. The quality analysis for the other templates has also been performed and is reported in the Supplementary Material, see Figures S1 and S2. Moreover, for each template and each point mutation we reported the quality value resulting from the QMEANBrane analysis, see Tables S3, S4, S5, S6, S7, S8 S9 and S10. It is worth noticing that the analysis in Fig. 6 is mainly relevant for the protein segments corresponding to the transmembrane regions, i.e. all the α-helices (see Fig. 1), because this is the critical region as previously discussed.

Figure 5. DCA analysis for templates 6A90 and 6J8J where green points are representing the contact map of the original template and the red points the corresponding DCA representations. (a) and (b) full sequence; (c) and (d): interval 1181–1851.

Figure 6a shows the predicted local similarity to the target structure, and highlights how the produced models have high quality within each domain area, and low quality in the inter-domains area, where the reliability of the models is significantly lower. However, this is completely acceptable as inter-domains region are formed by unstructured loops outside the membrane area (see Figs. 1 and 4), whereas all the considered mutations fall within the high-quality domain regions. As it will be further elaborated below, the low-quality loop regions enter in the network analysis without however jeopardizing the soundness of the results for the important high-quality domains region.

Additional insights can be obtained by performing a Ramachandran plot analysis, to assess whether the obtained structures obey all characteristic constraints provided by both steric hindrance and quantum chemistry. This is done in Fig. 6b, and shows that indeed 97.9% of the produced amino acids fall within a favorable region, 2% fall within a permitted region and only one amino acid falls in the forbidden region (the white region).
Armed by all these information, we now consider the issue of comparing the obtained different structures. As anticipated, our strategy hinged upon a mapping of each three-dimensional structure into a corresponding network (univocally identified by its topology) and then compare different networks with each other. We use RIN analysis to perform this mapping and both Vertex Histogram (VH) kernel and Weisfeiler–Lehman (WL) subtree kernel with 5 iterations, to compare them, as illustrated in section “Methods”. The similarity between all considered graphs can be visually assessed via the similarity matrix rendering reported in Fig. 7 for all three MOESM3, 6A90 and 6J8J templates. Each depicted similarity matrix of Fig. 7a–f (left panel) has rows and columns labeled according the mutation ids 0–84 of Fig. 4. Accordingly, 0–30 correspond to PAT mutations while 30–84 are NEUTRAL variants. Each cell ($i, j$) in a matrix shows the similarity value between the $i$-th and $j$-th RINs color-coded so that lighter colors correspond to RINs with high degree of similarity (from yellow to blue). Clearly the main diagonal shows always the lightest color, being the result of the comparison of a graph with itself.

Three main points emerge from the analysis of the results of Fig. 7a–f:

- In both the MOESM3 and 6A90 templates two well-defined clusters are clearly visible. The first one grouping together PAT mutations (ids 0–29) that have a high degree of similarity within each other and a low degree of similarity with NEUTRAL mutations (ids 30–84). Likewise NEUTRAL mutations are similar to each others but different from PAT ones.

- No such division in clusters is visible for the 6J8J template, where there is no clear distinction between PAT and NEUTRAL mutations. Although a few genetic variants appear to display a more marked difference, the majority show a high degree of similarity with each other. The reasons for this marked contrast with what has been observed for templates MOESM3 and 6A90 is unclear. One possible explanation stems from the observation that 6J8J is a human protein template representing the native state of the NaV1.7 sodium channel in closed state. The considered WT sequence of NaV1.7 and the 6J8J primary sequence are then nearly identical likely for all the 85 considered genetic variants (each variant differ from the WT for just one amino acid). Hence reconstructed models via homology modeling and energy minimization will also provide nearly identical RINs whose differences can hardly be captured by any pattern recognition algorithm such as that employed here. This is not the case of the two other templates MOESM3 and 6A90 where generated three dimensional structures via homology models might show a significant structural difference with one another so that energy minimization might drive them into different local minima. See the Supplementary Material, Figure S3.

- For all the three considered templates, the WL and VH kernels show a similar pattern. This means that the pairwise comparison of the nodes labels alone, as done by the VH kernel, is already sufficient to capture the main features. The WL kernel performs a more general comparison since also the subtree structure of each node is considered. Hence, while a pairwise comparison of the nodes labels is sufficient to discriminate the
and Neutral classes, the subtree structure and labels is also found to be coherent with this classification; We also examined the role of each interaction (H-bond, Van der Waals, and Ionic) separately and found the same pattern observed in the whole RINs. All these results can be found in the Supplementary Material, see Figures S4, S5.

Further support to the above results can be obtained using unsupervised machine learning techniques. To this aim we apply the Dominant Set (DS) algorithm to the above similarity matrices. The results for matrix (c) of the 6A90 template are reported the right panel of Fig. 7. The diagrams (g) and (h) are composed of 85 rows labeled with the ids of considered genetic variants and two columns. In the first column (labeled correct clusters) the first 31 PAT mutations are color-coded in black, and the remaining Neutral variants are color-coded in yellow.

Figure 7. [Left] Similarity matrices of the Weisfeiler–Lehman (5 iterations) (a)–(c)–(e) and Vertex Histogram (b)–(d)–(f) kernels applied to RINs resulting from MOESM3, 6A90 and 6J8I templates. [Right] Dominant set classification for the WL similarity matrix (c) of 6A90 template: first (g) and second (h) iterations.
The second column (labeled DS clusters) reports the classification according to the DS algorithm. Diagram (g) is the result of one iteration of the DS algorithm. Remarkably, the algorithm captures the correct classification of 96.8% of the pain related \textit{PAT} mutations, color-coded in black. Only one pain mutation (T1596I, id 30) is not correctly classified. This result is not surprising: looking at matrix (c) it is evident that mutation 30 is very different from the ones in the \textit{PAT} group and instead very similar to the ones in the \textit{NEUTRAL} group. Among the \textit{NEUTRAL} variants, color-coded in yellow, 13 variants are not classified correctly, 4 of which are human variants. An additional iteration of the DS algorithm within the variants originally classified in the black group further splits it in two clusters as shown in diagram (g): the first cluster, shown in blue, contains the 13 neutral variants plus the deleterious mutation (A863P, id 11) and the second one, shown in black, contains all the other deleterious mutations. Note that the lower degree of similarity of mutation (A863P, id 11) and the rest of the \textit{PAT} group is also visually evident from the similarity matrix (c) and can be due to a lower quality of the structure model.

Consistent findings are found by applying the average-linkage hierarchical clustering\textsuperscript{72} to the distance matrix derived from matrix (c), as presented in Fig. 8. Here, the leaves of the dendrogram are labeled with the ids of the considered genetic variants and the scale on the left shows the distance among clusters. Note that the \textit{PAT} mutations are clearly separated from the \textit{NEUTRAL} variants, as for the DS algorithm, but mutations 11 and 30 are further singled out, as suggested by the second iteration of the DS algorithm, diagram (g).

The clustering results for templates MOESM3 and 6J8J can be found in the Supplementary Material, see Figure S6.

We also applied supervised machine learning techniques in the form of the Support Vector Machine (SVM) method with cross-validation to determine its prediction ability. Cross-validation is a standard technique used to evaluate the ability of prediction models, see section “Methods” for a detailed description. It is particularly useful when a small dataset is available, as in our case. The accuracy of the predictions made for the three considered templates are shown in Table 1. As visible, the WL kernel predictions outperforms VH for all templates and the accuracy results are in line with the kernels results: the ability of the VH and WL kernels to discriminate
between pain related mutations and non pain ones is reflected also by the SVM learning method. In particular, the prediction accuracy is very good for templates MOESM3 and 6A90 and rather poor for 68J, in agreement with previous findings.

A word of caution is in order here. At first sight, it might appear striking the contradiction between the high sequence identity of the 68J hNav1.7 human template with the WT and the corresponding low accuracy of the prediction. However, this is in line with past work, as anticipated earlier, and can be ascribed to the difficulties of homology modelling when comparing sequences with too high identity (see e.g. Ref. 7), whereas performs better when comparing different species. A good example is provided by a recent study by Huang et al. 76 where the atomic structure of the rabbit voltage-gated calcium channel Cav1.1 provides a very good template for homology-based structural modeling of Nav channels notwithstanding a comparatively low value (21%) of sequence identity. This is more true within our computational pipeline because we are performing a pairwise comparison of structures to identify emerging global patterns. Additional human templates with lower sequence identity have also been recently proposed 76, but the corresponding reduced identity can there be ascribed to a significantly lower quality of the original deposited structure and to the much lower number of represented amino acids, as detailed in the Supplementary Material.

The general scenario emerging from previous analysis strongly suggests the existence of some critical mutations that lead to deleterious effect. This is an important issue with far reaching consequences in the case of human genome 78. Even at the level of globular proteins, it is well-known how single mutations of specific amino acids can change folds and functions in a very controlled way 79,80. However, this case appears to be more subtle. For instance, it is interesting to note that M1532I is a PAT genetic variant and M1532V is a NEUTRAL genetic variant notwithstanding the fact that Isoleucine (I) and Valine (V) are both hydrophobic amino acids with quite similar chemical structure. On this basis, both M1532I and M1532V were expected to behave similarly. Yet, there is significant evidence in the literature for the former to be a pathogenetic variant 76, whereas M1532V is reported in dbSNP 81, lacking information about the association to a clinical condition and about minor allele frequency.

The case of these two closely localized variants with a different predicted impact, points out the importance to implement this study with new variants in order to highlight even very subtle differences between residues, which can discriminate between high-impact and neutral variants. While this particular case is particularly striking, it turns out to be not the only one (see e.g. Huang et al. 76). In fact, we did not find any well-defined correlation between the chemical physical properties of the mutated amino acids and their putative pathogenic impact.

We can build on this idea in our case by presenting two different analyses that provide concurring results supporting the above predicted scenario.

The first analysis is prompted by the ability of the VH kernel to discern between PAT and NEUTRAL mutations, and hinges on the node frequency analysis reported in section “Methods”. On recalling that each node represents an amino acid of the protein sequence, labeled by its position in the sequence itself, and that the VH kernel is based on nodes labels comparison, we compute the frequency of each node label within the PAT and the NEUTRAL groups. Consider a node with label i (i.e. the amino acid in position i in the protein sequence), we then look to all RINs associated to PAT group—that can originate from any mutation of a specific node i = 1, . . . 2000 including i = 1, and count the number of times that that i is involved in the RINs of that group. A frequency 1 then means that i has non-covalent bonds in all mutations belonging to the PAT group, with a frequency 0 indicating that i has no bonds except the covalent ones. The same analysis can then clearly be performed for the NEUTRAL group, with similar interpretation.

Figure 9 shows the frequency of nodes along the protein sequence for the PAT and NEUTRAL groups. The range of the four domains are highlighted in orange along the sequence axis and the dashed lines identify the intervals exhibiting a frequency variation in one of the two groups and not in the other. Note that the majority of nodes have frequency one in both groups, meaning that all the corresponding RINs are very similar for large parts of the protein sequence. For the PAT group, this includes almost entirely the regions involved in the formation of the four domains, except for the terminal traits of D II and D IV. When not with frequency one, the PAT and NEUTRAL groups appear to be characterized by two substantially different patterns: the PAT group shows two broad bands located in the regions between 400–720 and 950–1150; while the NEUTRAL group is characterized by having three more picked bands: 400–610, 850–1000, and from 1700 onwards. Both these bands are highlighted with thick dotted lines in Fig. 9, red for PAT group, green for the NEUTRAL group.

The second analysis is based on the calculation of the entropy profile that can be obtained by the MSA calculation performed within the DCA approach, as outlined in section “Methods”. The main idea is that a low-entropy character is characteristic of conserved position, so we expect low entropy values for PAT mutations and high entropy

| Template | Kernel method | Prediction accuracy (%) |
|----------|---------------|-------------------------|
| MOESM3   | VH            | 63.5                    |
| 6A90     | VH            | 95.14                   |
| 68J      | VH            | 63.5                    |
| MOESM3   | WL            | 94.30                   |
| 6A90     | WL            | 97.64                   |
| 68J      | WL            | 63.5                    |

Table 1. Prediction accuracy of the SVM method for the three considered templates. The best results are highlighted in bold.
values for NEUTRAL mutations. Note that 'low' and 'high' are here defined with respect to the average value of the entropy in the mutated sites. Note that here, unlike the node frequency analysis, the entropy value refers to the \( i \)-th node (amino acid) that has been muted.

As Fig. 10 shows, only four PAT mutations occur in amino acids with higher-than-average entropy, so the overwhelming majority of these mutations occur in conserved regions with low entropy. By contrast, most of the

---

**Figure 9.** Relative frequency of nodes for pathogenic (PAT) and non pathogenic (NEUTRAL) mutations. Vertical dotted lines identify regions of different behaviour between the two groups, color-coded according to their specificity (see text).

**Figure 10.** Entropy profile of the full alignment: red dots highlight PAT mutations and blue dots the NEUTRAL ones. The horizontal line is the mean entropy value of the alignment in the mutated sites.
NEUTRAL mutations occur in higher-than-average entropy, indicating that those mutations occur in regions with high mutation propensity. Not surprisingly, PAT mutations with low entropy bear mutations of residues with indices located in DI (∼250), DII (∼800), DIII (∼1300), and DIV (∼1700) domains, in agreement with the intuitive fact that evolution has optimized sequences that try to avoid PAT mutations. When contrasted with the analysis of Fig. 9 we note that the second and the forth of the above regions are again located in the VSD domain, but the other two are not.

It would be interesting to pursue a better characterization of those mutations that do not follow this simple rule. However, a much larger dataset than that available at the present time would be needed to properly discuss the pathogenicity of some particular mutation. Ideally a detailed characterization of the effect of all possible mutations for that particular wild type residue would be desirable.

Conclusions

In short, we have implemented a computational pipeline to analyze the effects of a single residue mutation in human NaV1.7 channel. Our aim was to identify emerging patterns associated with gain-of-function mutations that lead to diseases by impairment of protein activity. The general workflow includes the following steps. Start with a specific template, perform a preliminary test via DCA analysis for its reliability, and then use homology modeling and energy minimization to find all different three-dimensional structures associated with each of the single amino acid variants; quality assessment of the obtained structures; a mapping of the three-dimensional structures into the corresponding topological network via RIN creation; machine learning of the similarity index of different RNs to identify emerging patterns.

We have applied this workflow to the specific case of different mutations of protein NaV1.7 that is involved in the propagation of nociception signals. We examined 31 gain-of-function mutations (PAT) that have been shown to cause a functional impairment of the channel demonstrated by cell electrophysiology assay and 21 benign or with uncertain significance variants from dbSNPs and 33 pseudo mutations identified among SCN9A shown to cause a functional impairment of the channel demonstrated by cell electrophysiology assay. We examined 31 gain-of-function mutations that have been identified in the propagation of nociception signals. We examined 31 gain-of-function mutations (PAT) that have been shown to cause a functional impairment of the channel demonstrated by cell electrophysiology assay and 21 benign or with uncertain significance variants from dbSNPs and 33 pseudo mutations identified among SCN9A shown to cause a functional impairment of the channel demonstrated by cell electrophysiology assay.

In the context of benign or with uncertain significance variants from dbSNPs and 33 pseudo mutations identified among SCN9A shown to cause a functional impairment of the channel demonstrated by cell electrophysiology assay and 21 benign or with uncertain significance variants from dbSNPs and 33 pseudo mutations identified among SCN9A shown to cause a functional impairment of the channel demonstrated by cell electrophysiology assay.

We have applied this workflow to the specific case of different mutations of protein NaV1.7 that is involved in the propagation of nociception signals. We examined 31 gain-of-function mutations (PAT) that have been shown to cause a functional impairment of the channel demonstrated by cell electrophysiology assay and 21 benign or with uncertain significance variants from dbSNPs and 33 pseudo mutations identified among SCN9A shown to cause a functional impairment of the channel demonstrated by cell electrophysiology assay. We examined 31 gain-of-function mutations (PAT) that have been shown to cause a functional impairment of the channel demonstrated by cell electrophysiology assay and 21 benign or with uncertain significance variants from dbSNPs and 33 pseudo mutations identified among SCN9A shown to cause a functional impairment of the channel demonstrated by cell electrophysiology assay.

In conclusion, this study represents a pilot application of prediction methods to separate deleterious from neutral genetic variants in Nav1.7 sodium channel, supporting the selection of candidates for cell electrophysiology assay.
Methods

Direct coupling analysis (DCA). Roughly speaking, the idea underlying DCA is that the mutation of any amino acid is constrained by the need of conserving protein function. In particular, a mutation of a residue involved in an important interaction with a second one can only be fixed during evolution if the new mutated residue has similar physical properties to the first one or if the second residue is also mutated in a concerted way. This mechanism gives rise to patterns of correlated mutations that can be detected in a multiple sequence alignment (MSA) of proteins of the same family and used to infer pair of residues that are most likely interacting in the tertiary structure. DCA has been proved to be able to correctly predict with good accuracy the contact map of proteins belonging to sufficiently large protein families. For this study we used an in-house code, which compute DCA using the pseudolikelihood approximation.

The inference was performed for only the regions of the alignment for which the fraction of gaps is less than 30%, which mainly correspond to the region from residue 1280 to residue 1840.

In particular, DCA analysis was obtained by performing an initial alignment comprising sequences Q15858 (Homo sapiens), Q62205 (Mus musculus), O08562 (Rattus norvegicus) and Q28644 (Oryctolagus cuniculus) from the Uniprot database. The alignment has been used as a seed for searching homologous sequences in Uniprot and Metaclust databases. An hidden markov model profile of the seed has been built with hmmbuild with default parameters. We searched the profile using hmmsearch with default parameters. We found 45447 matches on Uniprot and 104741 on Metaclust. We concatenated the sequences obtained from the two databases into a single MSA and filtered out all those sequences with more than 50% of gaps. The filtered MSA only contained 6637 sequences. In order to select a region of the protein for which we could obtain a lower fraction of gaps we selected from the full MSA only those columns with less than 30% of gaps, and we filtered out again those sequences with more than 70% gaps. This restricted MSA contains 19045 sequences.

Uniprot database has been downloaded on 2018/11/27, while we used Metaclust release 2018-06 for metagenomic data.

Homology modelling. Homology modeling relies on evolutionary related structures (templates) to generate a structural model of protein of interest (target). For this task we used Swiss-Model, which is widely known and used in the literature. It uses a sequence S and a template T (in PDB format) as input. The output generated is a tentative three-dimensional (3D) structure (in PDB format) associated with the sequence S. Swiss-model includes also an energy minimization step to resolve small structural distortions, unfavourable interactions or clashes introduced during the modelling process. Besides the 3D structure, also the quality of the obtained model is evaluated.

Energy minimization. The structures produced by homology modeling are often closer to the template than to their native structures. It is therefore necessary to perform an energy minimization step using dedicated tools. To this purpose we employed FG-MD: it uses a multiple templates approach to reshape the energy landscape from golf-course-like to funnel-like ones and drive the energy minimization closer to native state.

Quality assessment. QMEANBrane exploits the increasing availability of deposited high definition membrane protein structures to adapt knowledge-based methods to this class of proteins. It is known that the properties of membrane proteins are strongly influenced by their interaction with phospholipid tails, but a clear division into a membrane region and a soluble region does not adequately reflect the variation in molecular properties along the membrane axis. To capture these differences, QMEANBrane divides the study into three parts: an interface zone consisting of all those residues whose $C_n$ are at a distance of 5Å from the defined membrane plane; a membrane region enclosed by all those residues that are more than 5Å between the two planes; finally, a region of soluble protein consisting of the remaining amino acids.

Residue interaction networks (RIN). Given a protein structure, its corresponding RIN is generated by inspecting its atoms and looking for non-covalent bonds between them. We used RING 2.0 to perform this task. Given the PDB representation of a protein, the RING-2.0 algorithm generates the graph in two steps. The first identifies a list of residue-residue pairs eligible to undergo an interaction based merely on distance measurements. The second characterizes every contact by identifying the specific type of interaction, which can be Hydrogen bond, Van der Waals, Ionic bridges, π-π-Stacking, π-cation and Disulfide bridges.

RING 2.0 allows the RING generation according to various parameters that the user can select. All RINs produced in this study have been generated using the default options. In particular, for the Network policy parameter we used the Closest option, where all atoms of a residue-residue pair are considered to measure the distance; for the Interaction type we used the Multiple option, where RING reports multiple interactions per residue pair but only one interaction per interaction type; for the Distance threshold we used the Strict option, where each type of interaction has a specific threshold depending on its strength, as reported in Table 2.

Graph kernels. Graph kernels can be intuitively understood as functions measuring the similarity of pairs of graphs. We used two different kernels to compare RINs, the Vertex Histrogram (VH) and Weisfeiler–Lehman (WL) subtree kernels, implemented by the GraKel Python library. Both of them are based on node labels comparison, where node labels are set as the position of each node, i.e. amino acid, in the protein sequence. The essential difference between the two methods is that the VH method makes a pairwise comparison of the nodes labels only, whereas the WL method compares also the corresponding subtrees, hence providing a more com-
that are not strictly related to a particular partition of training and testing datasets. Even having a large dataset, using cross-validation gives more reliable results, because it produces a good result for accuracy it means that the problem is highly linearly separable and the model is not required to perturb the boundary. In fact, for each step, the support vectors are composed of different points. If cross-validation does not perform well on this task then cross-validation helps to divide two classes is maintained by a small subset of points, called support vectors. These points are the closest to the boundary, so adding more data does not affect predictions unless the new data is considerably closer to the boundary than the support vectors. Since we don’t have a large data set, the use of cross-validation helps to determine in sequence.

The Average Linkage is a well-known clustering method that creates clusters by proceeding in a bottom-up way. Given in input the distance matrix of a set of objects, it starts by assigning to each object its own cluster and then proceeds by joining at each step the two most similar clusters. In Average Linkage hierarchical clustering, the distance between two clusters is defined as the average distance between each object in one cluster to every object in the other cluster. We used the R Package implementation of the Average Linkage clustering and displayed the result as a dendrogram where all input objects are shown on the leaves and the scale represents the distance between clusters.

Clustering methods. We apply two different clustering methods: Dominant Set (DS) and Average-Linkage. The DS method tries to answer the problem of organizing a set of elements in clusters, in such a way that each group satisfies an internal homogeneity and expresses an external inhomogeneity between the groups. The algorithm behind this approach does not require any assumption underlying the data representation and does not require prior knowledge on the number of clusters to be obtained, as it is able to determine them in sequence. The Average Linkage is a well-known clustering method that creates clusters by proceeding in a bottom-up way. Given in input the distance matrix of a set of objects, it starts by assigning to each object its own cluster and then proceeds by joining at each step the two most similar clusters. In Average Linkage hierarchical clustering, the distance between two clusters is defined as the average distance between each object in one cluster to every object in the other cluster. We used the R Package implementation of the Average Linkage clustering and displayed the result as a dendrogram where all input objects are shown on the leaves and the scale represents the distance between clusters.

Support vector machine. In the basic configuration, Support Vector Machine (SVM) works with two classes and considers linearly separable problems, that is data points which can be separated by an hyperplane. Considering data as n-dimensional labeled points, SVM searches the hyperplane that separates points of different classes, maximizing the distance between the boundary and data points. For unsupervised learning, model’s prediction accuracy must be evaluated. To this aim we use the cross-validation technique, as implemented in the Scikit learn library.

Cross-validation is one of the standard techniques to check the prediction ability of the SVM supervised learning method. Roughly speaking, given a dataset of known data (the training set), cross validation divides the dataset into k partitions and, for each partition, one partition at a time is used as validation set, while the rest of the dataset remains as training set. At each iteration a value of prediction accuracy is measured and, at the end, the mean value of such measures will be the final prediction accuracy result. In our case the training set consists of (the representation of) the 85 considered genetic variants and we set k = 10, that is, we repeat the partition into validation set and training set 10 times.

Cross-validation is useful to avoid overfitting problems in accuracy evaluation. The boundary found by SVM to divide two classes is maintained by a small subset of points, called support vectors. These points are the closest to the boundary, so adding more data does not affect predictions unless the new data is considerably closer to the boundary than the support vectors. Since we don’t have a large data set, the use of cross-validation helps to perturb the boundary. In fact, for each step, the support vectors are composed of different points. If cross-validation produces a good result for accuracy it means that the problem is highly linearly separable and the model has a good level of generalization. Even having a large dataset, using cross-validation gives more reliable results, that are not strictly related to a particular partition of training and testing datasets.

Node frequency analysis. This analysis takes in consideration RINs nodes. We remind that each node represents an amino acid of the protein sequence, labeled by its position in the sequence itself. The idea is then

| Interaction type | Strict threshold (Å) |
|------------------|---------------------|
| Hydrogen bonds   | 3.5                 |
| Van der Waals    | 0.5                 |
| Ionic Bridge     | 4.0                 |
| π–π stacking     | 6.5                 |
| π–cation         | 5.0                 |

Table 2. Threshold distance values for the Strict option of RING 2.0.
to compute the frequency of each node with the aim of highlighting the nodes role in pathogenic and non pathogenic mutations.

In particular, we computed the frequency of each node label within the pathogenic mutations (PAT group) and, separately, within the normal variants (NEUTRAL group). For instance, a node label \( f_i(\alpha) \) (i.e. the amino acid \( \alpha \) at position \( i \) in the protein sequence) that is present in all RINs of the pathogenic mutations has frequency one for the PAT group. If the same node label \( i \) is also present in all RINs of the normal variants, its frequency is one as well for the NEUTRAL group. By comparing the two frequencies we could argue that \( i \) is not a sensible node position w.r.t. the PAT and NEUTRAL genetic variants.

Entropy analysis. The MSA calculated for DCA analysis can be also used to calculate the entropy profile of the protein, with the idea of shedding light on which amino acids, along the chain, are most likely to change during evolution. The entropy profile \( S_i \) of our MSA, has been calculated by first computing the frequency \( f_i(\alpha) \) of each non-gap symbols \( \alpha \) for a given position \( i \), and finally summing \( S_i = -\sum_{\alpha} f_i(\alpha) \ln(f_i(\alpha)) \) according with the Shannon formula. A high entropy is indicative of a high mutation propensity, and vice versa low entropy is characteristic of conserved positions.

References
1. Carterall, W. Structure and function of voltage-gated sodium channels at atomic resolution. Exp. Physiol. 99, 35–51 (2014).
2. De Lera Ruiz, M. & Kraus, R. Voltage-gated sodium channelic structure, function, pharmacology, and clinical indications. J. Med. Chem. 58, 7093–7118 (2015).
3. Emery, E., Luiz, A. & Wood, J. NaV1.7 and other voltage-gated sodium channels as drug targets for pain relief. Expert Opin. Ther. Targets 20, 975–983 (2016).
4. Ullman, S. Using neuroscience to develop artificial intelligence. Science 363, 692–693 (2019).
5. Waxman, S. G. et al. Sodium channel genes in pain-related disorders: phenotype–genotype associations and recommendations for clinical use. Lancet Neurol. 13, 1152–1160 (2014).
6. Kaperis, D. et al. Network topology of NaV1.7 mutations in sodium channel-related painful disorders. BMC Syst. Biol. 11, 28 (2017).
7. Lee, R. Protein model building using structural homology. Nature 356, 543–544 (1992).
8. RING 2.0 Web Server. http://protein.bio.unipd.it/ring/.
9. Piovesan, D., Minervini, G. & Tosatto, S. The RING 2.0 web server for high quality residue interaction networks. Nucleic Acids Res. 44, W367–W374 (2016).
10. Zhang, X. et al. Graph kernels. In Encyclopedia of Machine Learning (eds Sammut, C. & Webb, G. I.) 467–469 (Springer, Boston, 2011).
11. Rota Bulò, S. & Pelillo, M. Dominant-set clustering: a review. Encyclopedia of Machine Learning (Encyclopedia of Machine Learning) 44, 894–898 (2006).
12. Cox, J. J. et al. An scn9a channelopathy causes congenital inability to experience pain. Nature 444, 894–898 (2006).
13. Dre nth, J. P. et al. Mutations in sodium-channel gene scn9a cause a spectrum of human genetic pain disorders. J. Clin. Investig. 117, 3603–3609 (2007).
14. Geer, L. et al. The NCBI biosystems database. Nucleic Acids Res. 38, D492–D496 (2009).
15. Sunyaev, S. et al. Prediction of deleterious human alleles. Hum. Mol. Genet. 10, 591–597 (2001).
16. Care, M., Needham, C., Bulpitt, A. & Westhead, D. Deleterious SNP prediction: be mindful of your training data. Bioinformatics 23, 664–672 (2007).
17. Yue, P. & Moult, J. Identification and analysis of deleterious human snps. J. Mol. Biol. 356, 1263–1274 (2006).
18. Cox, J. J. et al. An scn9a channelopathy causes congenital inability to experience pain. Nature 444, 894–898 (2006).
19. Lee, R. Protein model building using structural homology. Nature 356, 543–544 (1992).
20. Wang, Y. et al. Prediction of deleterious human alleles. Hum. Mol. Genet. 10, 591–597 (2001).
21. Sunyaev, S. et al. Prediction of deleterious human alleles. Hum. Mol. Genet. 10, 591–597 (2001).
22. Care, M., Needham, C., Bulpitt, A. & Westhead, D. Deleterious SNP prediction: be mindful of your training data. Bioinformatics 23, 664–672 (2007).
23. Adzhubei, I. et al. A method and server for predicting damaging missense mutations. Nat. Methods 7, 248–249 (2010).
24. Lee, M.-J. et al. Identification and analysis of deleterious human snps. J. Mol. Biol. 356, 1263–1274 (2006).
25. Care, M., Needham, C., Bulpitt, A. & Westhead, D. Deleterious SNP prediction: be mindful of your training data. Bioinformatics 23, 664–672 (2007).
26. Adzhubei, I. et al. A method and server for predicting damaging missense mutations. Nat. Methods 7, 248–249 (2010).
27. Yu, P. & Moult, J. Identification and analysis of deleterious human snps. J. Mol. Biol. 356, 1263–1274 (2006).
28. Care, M., Needham, C., Bulpitt, A. & Westhead, D. Deleterious SNP prediction: be mindful of your training data. Bioinformatics 23, 664–672 (2007).
29. Adzhubei, I. et al. A method and server for predicting damaging missense mutations. Nat. Methods 7, 248–249 (2010).
30. Lee, M.-J. et al. Identification and analysis of deleterious human snps. J. Mol. Biol. 356, 1263–1274 (2006).
31. Care, M., Needham, C., Bulpitt, A. & Westhead, D. Deleterious SNP prediction: be mindful of your training data. Bioinformatics 23, 664–672 (2007).
32. Adzhubei, I. et al. A method and server for predicting damaging missense mutations. Nat. Methods 7, 248–249 (2010).
33. Lee, M.-J. et al. Identification and analysis of deleterious human snps. J. Mol. Biol. 356, 1263–1274 (2006).
34. Care, M., Needham, C., Bulpitt, A. & Westhead, D. Deleterious SNP prediction: be mindful of your training data. Bioinformatics 23, 664–672 (2007).
35. Adzhubei, I. et al. A method and server for predicting damaging missense mutations. Nat. Methods 7, 248–249 (2010).
36. Lee, M.-J. et al. Identification and analysis of deleterious human snps. J. Mol. Biol. 356, 1263–1274 (2006).
37. Care, M., Needham, C., Bulpitt, A. & Westhead, D. Deleterious SNP prediction: be mindful of your training data. Bioinformatics 23, 664–672 (2007).
38. Adzhubei, I. et al. A method and server for predicting damaging missense mutations. Nat. Methods 7, 248–249 (2010).
39. Lee, M.-J. et al. Identification and analysis of deleterious human snps. J. Mol. Biol. 356, 1263–1274 (2006).
40. Care, M., Needham, C., Bulpitt, A. & Westhead, D. Deleterious SNP prediction: be mindful of your training data. Bioinformatics 23, 664–672 (2007).
41. Adzhubei, I. et al. A method and server for predicting damaging missense mutations. Nat. Methods 7, 248–249 (2010).
42. Lee, M.-J. et al. Identification and analysis of deleterious human snps. J. Mol. Biol. 356, 1263–1274 (2006).
43. Care, M., Needham, C., Bulpitt, A. & Westhead, D. Deleterious SNP prediction: be mindful of your training data. Bioinformatics 23, 664–672 (2007).
44. Adzhubei, I. et al. A method and server for predicting damaging missense mutations. Nat. Methods 7, 248–249 (2010).
45. Lee, M.-J. et al. Identification and analysis of deleterious human snps. J. Mol. Biol. 356, 1263–1274 (2006).
46. Care, M., Needham, C., Bulpitt, A. & Westhead, D. Deleterious SNP prediction: be mindful of your training data. Bioinformatics 23, 664–672 (2007).
34. Cummins, T. R., Dib-Hajj, S. D. & Waxman, S. G. Electrophysiological properties of mutant nav1.7 sodium channels in a painful inherited neuropathy. J. Neurosci. 24, 8232–8236 (2004).

35. Yang, Y. et al. Mutations in SCN9A, encoding a sodium channel alpha subunit, in patients with primary erythralgia. J. Med. Genet. 41, 171–174 (2004).

36. Theile, J. W., Jarecki, B. W., Piekarz, A. D. & Cummins, T. R. Nav1.7 mutations associated with paroxysmal extreme pain disorder, but not erythralgia, enhance nav4/4 peptide-mediated resurgent sodium currents. J. Physiol. 589, 597–608 (2011).

37. Chen, C. et al. Early- and late-onset inherited erythralgia: genotype-phenotype correlation. Brain 132, 1711–1722 (2009).

38. Han, C. et al. Sporadic onset of erythralgia: a gain-of-function mutation in nav1.7. Ann. Neurol. 59, 553–558 (2006).

39. Cheng, X. et al. Deletion mutation of sodium channel NaV1.7 in inherited erythralgia: enhanced slow inactivation modifies dorsal root ganglion neuron hyperexcitability. Brain 134, 1972–1986 (2011).

40. Harty, T. P. et al. Nav1.7 mutant a683p in erythralgia: effects of altered activation and steady-state inactivation on excitability of nociceptive dorsal root ganglion neurons. J. Neurosci. 26, 12566–12575 (2006).

41. Choi, J.-S. et al. Mexiteline-responsive erythralgia due to a new nav1.7 mutation showing use-dependent current fall-off. Exp. Neurol. 216, 383–389 (2009).

42. Cheng, X. et al. Mutations at opposite ends of the diii/s4–s5 linker of sodium channel na v 1.7 produce distinct pain disorders. Mol. Pain 6, 24–24 (2010).

43. Coordinates, N. R. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 44, D7–D19 (2015).

44. Estacion, M. et al. A new nav1.7 mutation in an erythralgia patient. Biochem. Biophys. Res. Commun. 432, 99–104 (2013).

45. Dib-Hajj, S. D. et al. Gain-of-function mutation in Nav1.7 in familial erythralgia induces bursting of sensory neurons. Brain 128, 1847–1854 (2005).

46. Greig, B. et al. Novel mutations mapping to the fourth sodium channel domain of nav1.7 result in variable clinical manifestations of primary erythralgia. Neuron Mol. Med. (2013).

47. Fertleman, C. R. et al. SCN9A mutations in paroxysmal extreme pain disorder: allelic variants underlie distinct channel defects and phenotypes. Neuron 52, 767–774 (2006).

48. Jarecki, B. W., Sheets, P. L., Jackson, J. O. II & Cummins, T. R. Paroxysmal extreme pain disorder mutations within the d3/s4–s5 linker of nav1.7 cause moderate destabilization of fast inactivation. J. Physiol. 586, 4137–4153 (2008).

49. Choi, J.-S. et al. Paroxysmal extreme pain disorder: a molecular lesion of peripheral neurons. Nat. Rev. Neurol. 7, 51–55 (2011).

50. Dib-Hajj, S. D. et al. Paroxysmal extreme pain disorder m1627k mutation in human nav1.7 renders drg neurons hyperexcitable. Mol. Paink (2008).

51. Theile, J. W. & Cummins, T. R. Inhibition of Nav9/4 peptide-mediated resurgent sodium currents in Nav1.7 channels by carba-mazepine, riluzole, and anandamide. Mol. Pharmacol. 80, 724–734 (2011).

52. Estacion, M. et al. Nav1.7 gain-of-function mutations as a continuum: A1632e displays physiological changes associated with erythralgia and paroxysmal extreme pain disorder mutations and produces symptoms of both disorders. J. Neurosci. 28, 11079–11088 (2008).

53. Faber, C. G. et al. Gain of function nav1.7 mutations in idiopathic small fiber neuropathy. Ann. Neurol. 71, 26–39 (2012).

54. Estacion, M. et al. Intra- and interfamily phenotypic diversity in small-fiber pain syndromes associated with a gain-of-function variant of nav1.7. Mol. Painhttps://doi.org/10.1186/1744-8069-7-92 (2011).

55. Han, C. et al. Nav1.7-related small fiber neuropathy. Neurology 78, 1635–1643 (2012).

56. Hoehm, J. et al. Small nerve fibres, small hands and small feet: a new syndrome of pain, dysautonomia and acromesomelia in a kindred with a novel Nav1.7 mutation. Brain 135, 345–358 (2012).

57. Bennett, D. L., Clark, A. J., Huang, J., Waxman, S. G. & Dib-Hajj, S. D. The role of voltage-gated sodium channels in pain signaling. Physiol. Rev. 99, 1079–1153 (2019).

58. Blesneac, I. et al. Rare nav1.7 variants associated with painful diabetic peripheral neuropathy. Pain 159 (2018).

59. Stevers, F. & Higgins, D. Clustal omega. Curr. Protoc. Bioinform. 48, 3.13.1–3.13.16 (2014).

60. Payandeh, J., Scheuer, T., Zheng, N. & Catterall, W. A. The crystal structure of a voltage-gated sodium channel. Nature 475, 353–358 (2011).

61. Shen, H. et al. Structural basis for the modulation of voltage-gated sodium channels by animal toxins. Science 362, eaao2596 (2018).

62. Shen, H., Liu, D., Wu, K., Lei, J. & Yan, N. Structures of human nav1.7 channel in complex with auxiliary subunits and animal toxins. Science 363, 1303–1308 (2019).

63. Morcos, F. et al. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proc. Natl. Acad. Sci. 108, E1293–E1301 (2011).

64. Waterhouse, A. et al. Swiss-model: homology modelling of protein structures and complexes. Nucleic acids Res. 46, W296–W303 (2018).

65. Swiss Institute of Bioinformatics. https://swissmodel.expasy.org/.

66. Zhang, Y., Liang & Zhang, Y. Atomic-level protein structure refinement using fragment-guided molecular dynamics conformation sampling. Structure 19 (2011).

67. Studer, G., Bissini, M. & Schwede, T. Assessing the local structural quality of transmembrane protein models using statistical potentials (qmeanbrane). Bioinformatics30 (2014).

68. Lovell, S. et al. Structure validation by Ca geometry: φ, ψ and cβ deviation. Proteins Struct. Funct. Bioinform. 50, 437–450 (2003).

69. Huang, W., Liu, M., Yan, S. F. & Yan, N. Structure-based assessment of disease related mutations in human voltage-gated sodium channels. Protein Cell 8, 401–438 (2017).

70. Evgeniou, T. & Pontil, M. Support vector machines: theory and applications. In Machine Learning and Its Applications Vol. 2049 (eds Goos, G. et al.) 249–257 (Springer, Berlin, 2001).

71. Siglids, G. et al. Grakel: a graph kernel library in Python. arXiv:abs/1806.02193 (2018).

72. Everitt, B., Landau, S., Lee, M. & Stahl, D. Cluster Analysis 5th edn. (Wiley, New York, 2010).

73. Gunter, S., Bierne, N. & Weissenbach, J. Structural basis of the voltage-gated calcium channel cav1.1 complex. Science (2015).

74. Wu, J. et al. Structure of the voltage-gated calcium channel cav1.1 complex. Science 348, 191–196 (2015).

75. Xu, H. et al. Structural basis of the paroxysmal extreme pain disorder inhibiting by a gating-modifier spider toxin. Cell 176, 702–715 (2019).

76. Haddad, Y., Adam, V. & Heger, Z. Ten. quick tips for homology modeling of high-resolution protein 3d structures. PLoS Comput. Biol. 16, e1007449 (2020).

77. Lohmueller, K. E. et al. Proportionally more deleterious mutation in European than in African populations. Nature 451, 994–997 (2008).

78. Alexander, P. A., He, Y., Chen, Y., Orban, J. & Bryan, P. N. The design and characterization of two proteins with 88% sequence identity but different structure and function. Proc. Natl. Acad. Sci. 104, 11963–11968 (2007).

79. Alexander, P. A., He, Y., Chen, Y., Orban, J. & Bryan, P. N. A minimal sequence code for switching protein structure and function. Proc. Natl. Acad. Sci. 105, 21149–21154 (2008).

80. Faber, C. G. et al. Gain of function nav1.7 mutations in idiopathic small fiber neuropathy. Ann. Neurol. 71, 26–39 (2012).

81. Sherry, S. T. et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 29, 308–311 (2001).
83. Han, C. et al. Human nav1.8: enhanced persistent and ramp currents contribute to distinct firing properties of human drg neurons. *J. Neurophysiol.* 113, 3172–3185 (2015).
84. Dca-epfl. https://gitlab.com/LBS-EPFL/code/lbsDCA/-/tree/v1.0.
85. Malinverni, D. & Barducci, A. Coevolutionary Analysis of Protein Sequences for Molecular Modeling Vol. 2022, 379–397 (Humana Press, New York, 2019).
86. UniProt. https://www.uniprot.org/.
87. Steinegger, M. & Söding, J. Clustering huge protein sequence sets in linear time. *Nat. Commun.* 9, 2542. https://doi.org/10.1038/s41467-018-04964-5 (2018).
88. Nikolentzos, G., Siglidi, G. & Vazirgiannis, M. Graph kernels: a survey. arXiv e-print (2019).
89. Shervashidze, N., Schweitzer, P., Leeuwen, E.J., Mehlhorn, K. & Borgwardt, K. M. Weisfeiler–Lehman graph kernels. *J. Mach. Learn. Res.* 12, 2539–2561 (2011).
90. Sugiyama, M. & Borgwardt, K. Halting in random walk kernels. In Advances in Neural Information Processing Systems Vol. 28 (eds Cortes, C. et al.) 1639–1647 (Curran Associates, New York, 2015).
91. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing (2013).

**Acknowledgements**

We thank Dimos Kapetis for useful discussions. The use of the SCSCF multiprocessor cluster at the Università Ca' Foscari Venezia is gratefully acknowledged. The work was supported by MIUR PRIN-COFIN2017 Soft Adaptive Networks grant 2017Z55KCW (A.G).

**Author contributions**

M.S., G.L. and A.G. conceived the research. A.T., G.C., S.Z., and M.S. performed the research. A.T, A.G., and M.S. analyzed the data and wrote the paper. All authors discussed the results and reviewed the manuscript.

**Competing interests**

The authors declare no competing interests.

**Additional information**

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-020-74591-y.

**Correspondence** and requests for materials should be addressed to M.S.

**Reprints and permissions information** is available at www.nature.com/reprints.

**Publisher’s note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

---

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2020