PLOS ONE

Predicting hotspots for disease-causing single nucleotide variants using sequences-based coevolution, network analysis, and machine learning

--Manuscript Draft--

Manuscript Number:	PONE-D-23-33350R1
Article Type:	Research Article
Full Title:	Predicting hotspots for disease-causing single nucleotide variants using sequences-based coevolution, network analysis, and machine learning
Short Title:	Predicting nSNVs hotspots with sequences-based machine learning
Corresponding Author:	Wenjun Zheng University at Buffalo, The State University of New York Buffalo, NY UNITED STATES
Keywords:	Centrality, Coevolution, Disease mutations, Machine learning, Protein residue contact network, Single nucleotide variant
Abstract:	To enable personalized medicine, it is important yet highly challenging to accurately predict disease-causing mutations in target proteins at high throughput. Previous computational methods have been developed using evolutionary information in combination with various biochemical and structural features of protein residues to discriminate neutral vs. deleterious mutations. However, the power of these methods is often limited because they either assume known protein structures or treat residues independently without fully considering their global interactions. To address the above limitations, we build upon recent progress in machine learning, network analysis, and protein language models, and develop a sequences-based variant site prediction workflow based on the protein residue contact networks: 1. We employ and integrate various methods of building protein residue networks using state-of-the-art coevolution analysis tools (RaptorX, DeepMetaPSICOV, and SPOT-Contact) powered by deep learning. 2. We use machine learning algorithms (Random Forest, Gradient Boosting, and Extreme Gradient Boosting) to optimally combine 20 network centrality scores to jointly predict key residues as hot spots for disease mutations. 3. Using a dataset of 107 proteins rich in disease mutations, we rigorously evaluate the network scores individually and collectively (via machine learning). This work supports a promising strategy of combining an ensemble of network scores based on different coevolution analysis methods (and optionally predictive scores from other methods) via machine learning to predict candidate sites of disease mutations, which will inform downstream applications of disease diagnosis and targeted drug design.
Order of Authors:	Wenjun Zheng
Opposed Reviewers:	
Response to Reviewers:	see attached file.
Additional Information:	
Question	**Response**
Financial Disclosure Enter a financial disclosure statement that describes the sources of funding for the work included in this submission. Review the submission guidelines for detailed requirements. View published research articles from *PLOS ONE* for specific examples.	Yes

This statement is required for submission and **will appear in the published article** if the submission is accepted. Please make sure it is accurate. **Funded studies** Enter a statement with the following details: • Initials of the authors who received each award • Grant numbers awarded to each author • The full name of each funder • URL of each funder website • Did the sponsors or funders play any role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript? Did you receive funding for this work?	
Please add funding details. as follow-up to "**Financial Disclosure** Enter a financial disclosure statement that describes the sources of funding for the work included in this submission. Review the submission guidelines for detailed requirements. View published research articles from *PLOS ONE* for specific examples. This statement is required for submission and **will appear in the published article** if the submission is accepted. Please make sure it is accurate. **Funded studies** Enter a statement with the following details: • Initials of the authors who received each award • Grant numbers awarded to each author • The full name of each funder • URL of each funder website • Did the sponsors or funders play any role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript? Did you receive funding for this work?"	This study is funded by a grant from NIH.

Please select the country of your main research funder (please select carefully as in some cases this is used in fee calculation). as follow-up to "**Financial Disclosure** Enter a financial disclosure statement that describes the sources of funding for the work included in this submission. Review the submission guidelines for detailed requirements. View published research articles from *PLOS ONE* for specific examples. This statement is required for submission and **will appear in the published article** if the submission is accepted. Please make sure it is accurate. **Funded studies** Enter a statement with the following details: • Initials of the authors who received each award • Grant numbers awarded to each author • The full name of each funder • URL of each funder website • Did the sponsors or funders play any role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript? Did you receive funding for this work?"	UNITED STATES - US
Competing Interests Use the instructions below to enter a competing interest statement for this submission. On behalf of all authors, disclose any competing interests that could be perceived to bias this work—acknowledging all financial support and any other relevant financial or non-financial competing interests. This statement is required for submission and **will appear in the published article** if the submission is accepted. Please make sure it is accurate and that any funding	The authors have declared that no competing interests exist.

sources listed in your Funding Information later in the submission form are also declared in your Financial Disclosure statement. View published research articles from *PLOS ONE* for specific examples. **NO authors have competing interests** Enter: *The authors have declared that no competing interests exist.* **Authors with competing interests** Enter competing interest details beginning with this statement: *I have read the journal's policy and the authors of this manuscript have the following competing interests: [insert competing interests here]* * typeset	
Ethics Statement Enter an ethics statement for this submission. This statement is required if the study involved: • Human participants • Human specimens or tissue • Vertebrate animals or cephalopods • Vertebrate embryos or tissues • Field research Write "N/A" if the submission does not require an ethics statement. General guidance is provided below. Consult the submission guidelines for detailed instructions. **Make sure that all information entered here is included in the Methods section of the manuscript.**	N/A

Format for specific study types **Human Subject Research (involving human participants and/or tissue)** • Give the name of the institutional review board or ethics committee that approved the study • Include the approval number and/or a statement indicating approval of this research • Indicate the form of consent obtained (written/oral) or the reason that consent was not obtained (e.g. the data were analyzed anonymously) **Animal Research (involving vertebrate animals, embryos or tissues)** • Provide the name of the Institutional Animal Care and Use Committee (IACUC) or other relevant ethics board that reviewed the study protocol, and indicate whether they approved this research or granted a formal waiver of ethical approval • Include an approval number if one was obtained • If the study involved *non-human primates*, add *additional details* about animal welfare and steps taken to ameliorate suffering • If anesthesia, euthanasia, or any kind of animal sacrifice is part of the study, include briefly which substances and/or methods were applied **Field Research** Include the following details if this study involves the collection of plant, animal, or other materials from a natural setting: • Field permit number • Name of the institution or relevant body that granted permission	
Data Availability Authors are required to make all data underlying the findings described fully available, without restriction, and from the time of publication. PLOS allows rare exceptions to address legal and ethical concerns. See the PLOS Data Policy and FAQ for detailed information.	Yes - all data are fully available without restriction

A Data Availability Statement describing where the data can be found is required at submission. Your answers to this question constitute the Data Availability Statement and **will be published in the article**, if accepted. **Important:** Stating 'data available on request from the author' is not sufficient. If your data are only available upon request, select 'No' for the first question and explain your exceptional situation in the text box. Do the authors confirm that all data underlying the findings described in their manuscript are fully available without restriction?	
Describe where the data may be found in full sentences. If you are copying our sample text, replace any instances of XXX with the appropriate details. • If the data are **held or will be held in a public repository**, include URLs, accession numbers or DOIs. If this information will only be available after acceptance, indicate this by ticking the box below. For example: *All XXX files are available from the XXX database (accession number(s) XXX, XXX.*). • If the data are all contained **within the manuscript and/or Supporting Information files**, enter the following: *All relevant data are within the manuscript and its Supporting Information files.* • If neither of these applies but you are able to provide **details of access elsewhere**, with or without limitations, please do so. For example: *Data cannot be shared publicly because of [XXX]. Data are available from the XXX Institutional Data Access / Ethics Committee (contact via XXX) for researchers who meet the criteria for access to confidential data.* *The data underlying the results presented in the study are available from (include the name of the third party*	All relevant data and code are available at https://simtk.org/projects/hotspots

and contact information or URL). • This text is appropriate if the data are owned by a third party and authors do not have permission to share the data. * typeset	
Additional data availability information:	

PONE-D-23-33350
Predicting hotspots for disease-causing single nucleotide variants using sequences-based coevolution, network analysis, and machine learning

Dear Editor,

Thank you for reviewing my manuscript. Following your and the reviewers' comments, I have revised the manuscript accordingly to address the points raised during the review process.

In this resubmission, I have included the following items as requested:

- A rebuttal letter (labeled 'Response to Reviewers') that responds to each point raised by the academic editor and reviewer(s).
- A marked-up copy of the manuscript that highlights changes made to the original version. It is labeled 'Revised Manuscript with Track Changes'.
- An unmarked version of the revised paper without tracked changes. It is labeled 'Manuscript'.

Figure files were uploaded separately from the manuscript (Fig1.tif etc).

Sincerely,

Wenjun Zheng, PhD

Predicting hotspots for disease-causing single nucleotide variants using sequences-based coevolution, network analysis, and machine learning

Short title: Predicting nSNVs hotspots with sequences-based machine learning

Wenjun Zheng [1*]

[1] Department of Physics, State University of New York at Buffalo, NY 14260, United States of America

* Corresponding author

E-mail: wjzheng@buffalo.edu (WZ)

Keywords: Centrality, Coevolution, Disease mutations, Machine learning, Protein residue contact network, Single nucleotide variant

Abstract

To enable personalized medicine, it is important yet highly challenging to accurately predict disease-causing mutations in target proteins at high throughput. Previous computational methods have been developed using evolutionary information in combination with various biochemical and structural features of protein residues to discriminate neutral vs. deleterious mutations. However, the power of these methods is often limited because they either assume known protein structures or treat residues independently without fully considering their global interactions. To address the above limitations, we build upon recent progress in machine learning, network analysis, and protein language models, and develop a sequences-based variant site prediction workflow based on the protein residue contact networks: 1. We employ and integrate various methods of building protein residue networks using state-of-the-art coevolution analysis tools (RaptorX, DeepMetaPSICOV, and SPOT-Contact) powered by deep learning. 2. We use machine learning algorithms (Random Forest, Gradient Boosting, and Extreme Gradient Boosting) to optimally combine 20 network centrality scores to jointly predict key residues as hot spots for disease mutations. 3. Using a dataset of 107 proteins rich in disease mutations, we rigorously evaluate the network scores individually and collectively (via machine learning). This work supports a promising strategy of combining an ensemble of network scores based on different coevolution analysis methods (and optionally predictive scores from other methods) via machine learning to predict candidate sites of disease mutations, which will inform downstream applications of disease diagnosis and targeted drug design.

Introduction

The holy grail of structural biology is to solve high-resolution biomolecular structures at the genomic scale to inform mechanistic studies of their functions. Thanks to recent revolutions in computational structural biology (accurate protein structure prediction by AlphaFold [1] and RoseTTAFold [2]), it is now feasible to predict native structures for many proteins given their sequences (with some caveats, see [3]), thus practically solving the protein folding problem [4]. However, it remains challenging to predict dynamic structural ensembles [5] and mutation-induced effects [6] to meet the demand of mechanistic studies of protein functions and dysfunctions. While the public databases of protein sequences and variations increase rapidly owning to genomic/metagenomic sequencing efforts (the MetaClust database contains about 1.6 billion protein sequence fragments [7]), the growth of experimental protein structures [8] and predicted structures remains to catch up (the AlphaFold database contains over 200 million predicted structures [9]). Such sequences-structures gap has motivated the development of new computational tools that make functional sense of protein sequences without directly using structural information (for example, by using deep learning to train large protein language models [10]). Recently, AlphaMissense attained state of the art prediction of missense variant pathogenicity by adapting AlphaFold fine-tuned on human and primate variant population frequency databases [11].

A major interest in personalized medicine is in understanding novel genetic variations through genotype-phenotype association studies in relation to diseases. Particularly, a rapidly growing number of non-synonymous single nucleotide variants (nSNVs) have been uncovered in protein coding regions that can adversely impact protein function and cause diseases [12].

Various computational methods were developed using evolutionary conservation and phylogeny in combination with biochemical and structural properties of amino acids to discriminate neutral vs. deleterious nSNVs [13-22]. Protein structural dynamics has also proven useful in discovering functionally important residues [23,24] which could constitute hot spots for disease-causing nSNVs [25,26]. However, the requirement of 3D structures has limited the number of nSNVs that can be analyzed by existing structure-based computational tools, although such constraint has been significantly alleviated by recent progress in protein structure prediction [27].

As alternatives to structure-based methods, sequences-based coevolution analysis has become increasingly powerful in predicting structural couplings between pairs of contacting residues [28-31] , thanks to the development of direct coupling methods that can overcome the confounding indirect coupling effects [29,32,33] . In principle, coevolving pairs of residues can be identified from a sufficiently large multiple sequence alignment, allowing the prediction of close spatial proximity in the native structures. Boosted by deep learning and other algorithmic developments, this coevolution analysis has led to accurate prediction of residue contacts which make *de novo* protein structure prediction possible [28] . Furthermore, coevolution analysis (enhanced by deep learning) has also been used to study various aspects of protein functional interactions such as allostery [34] . For example, RaptorX uses an ultra-deep neural network combining coevolution information with sequence conservation information to infer 3D contacts with higher accuracy than previous methods [35,36]. DeepMetaPSICOV [37] combines the input feature sets used by earlier methods (MetaPSICOV [38] and DeepCov [39]) as input to a deep, fully convolutional residual neural network. SPOT-Contact predicts protein contact maps by stacking residual convolutional networks with two-dimensional residual bidirectional recurrent LSTM networks, and using both one-dimensional sequence-based and two-dimensional

evolutionary coupling based information [40]. These three state-of-the-art coevolution analysis methods are employed in this study to construct protein residue contact maps for network analysis (see below).

Another line of protein research is based on the treatment of a protein as a network where amino acid residues are nodes and their bonded/non-bonded interactions form edges [41]. Such models can be readily built upon 3D native structures so that a whole suite of network analysis tools (see https://networkx.org/) can be applied. For example, Amitai et al [42] used network analysis of protein structures (using closeness centrality) to identify functional residues. Going beyond network analysis, deep-learning-based study of protein graph neural networks is an active area of research [43].

In a recent paper, Butler et al [44] proposed a sequence-based Gaussian network model (Seq-GNM) to calculate the dynamic profile of a protein without a 3D structure. They used coevolution analysis to build a network model which connects residues predicted to be in contact via evolutionary couplings. Their work built on previous studies that shown crystallographic B-factors are useful in predicting the impact of nSNVs on protein function [45,46] : rigid sites with low B-factors are more susceptible to destabilizing nSNVs than flexible sites with high B-factors. Indeed, existing computational tools to diagnose neutral and deleterious nSNVs (such as PolyPhen-2 [47]) use crystallographic B-factors along with other evolutionary and structural features. More specifically, Butler et al used Seq-GNM to compute B-factors for protein residues, and they found that deleterious nSNVs are overabundant at low B-factor sites, while neutral nSNVs are overabundant at high B-factor sites. Mechanistically, low B-factors may indicate that a site is crucial for maintaining structural stability and/or modulating functional motions (as a hinge) and thus susceptible to mutations. In contrast, high B-factors are associated

with flexible regions with minimal interactions, which are thus more robust to mutations. Based on these observations, they proposed that the sequences-based predicted B-factors can discriminate between deleterious and neutral nSNVs without structural information.

Inspired by the above study and recent progress in machine learning, network analysis, and protein language models, we further develop the sequences-based protein residue network analysis in the following directions: 1. We build protein residue networks using three different coevolution analysis tools (RaptorX, DeepMetaPSICOV, and SPOT-Contact) as enabled by deep learning. 2. We exploit three machine learning algorithms (Random Forest, Gradient Boosting, and Extreme Gradient Boosting) to optimally combine 20 distinct network node centrality scores as calculated from the contact probability matrices to predict hot spot residues for disease mutations. 3. Based on a dataset of 107 proteins with known deleterious/neutral mutations, we evaluate our sequences-based network scores both individually and in combination, and then compare with alternative structures-based network scores and a physics force field based method. By optimally combing three coevolution analysis methods and the resulting 20 network scores by machine learning, we are able to discriminate deleterious and neutral mutation sites accurately (AUC of ROC ~ 0.84), which is on par with structure-based network scores (AUC ~ 0.83). Furthermore, by combining our method with a state-of-the-art predictor of the functional effects of sequence variation based on large protein language models (ESM [48]), we have significantly improved the prediction of disease variant sites (AUC ~ 0.89).

In the following sections, we first describe the detailed methodology in the order of the proposed workflow, then we report the results of evaluation of our network-based scores both individually and collectively (via machine learning), finally we discuss specific case studies of four proteins to illustrate the usage of our method.

Materials and methods

Here is a summary of the workflow of our sequences-based method:

a. Collect datasets of protein sequences and variants (see Section 1)
b. Run co-evolution analysis of a given target protein sequence to build a residue contact map P (see Section 2)
c. Use NetworkX to calculate node centrality scores based on P (see Section 3)
d. Use sequence-based GNM to calculate additional node scores (see Section 4)
e. (optional) Use protein language model (ESM) to predict variant importance (see Section 5)
f. (optional) Use AlphaFold and FoldX to predict variant importance (see Section 6 and 7)
g. Use machine learning to optimally combine the above scores for classifying deleterious vs neutral variant sites (see Section 8)

1. Datasets of protein sequences and variants

A dataset of 107 protein sequences with ≤500 residues and ≥20 annotated deleterious/neutral variants were collected from the HumVar database [47] (sources: humvar-2011_12.deleterious.pph.input and humvar-2011_12.neutral.pph.input from ftp://genetics.bwh.harvard.edu/pph2/training/training-2.2.2.tar.gz). Their UniProt ids and sequences can be accessed at https://simtk.org/projects/hotspots. This diverse dataset contains 97 proteins with their pairwise sequence identity < 30%.

The HumVar dataset consists of 13,032 human disease-causing mutations from UniProt and 8,946 human nonsynonymous single-nucleotide polymorphisms (nsSNPs) without annotated involvement in disease. This dataset was previously used to train and test PolyPhen-2 [47] for predicting damaging effects of missense mutations, and was used by Butler et al [44] in benchmarking their seq-GNM method for predicting deleterious/neutral nSNVs.

Since this dataset is highly imbalanced (there are 4040 deleterious mutation sites but only 120 neutral mutation sites) [49], we have added 3403 additional neutral sites with very low conservation scores (i.e. grade ≤ 2 as assessed by the ConSurf program [50]). Our objective is to train and test a binary classifier of residues in these proteins as deleterious or neutral. To this end, we split 107 proteins into training and testing sets (with 79 and 28 proteins, respectively), and perform evaluations based on the testing set. The main metric of evaluation is the ROC curves and associated area under the curve (AUC). AUC is a standard metric for evaluating binary classifiers based on the ROC curve of sensitivity and specificity. The ROC curves are also used in other computational papers for variant prediction (see [47]).

2. Sequences-based coevolution analysis and protein contact map construction

We perform coevolution analysis using three state-of-the-art methods: the RaptorX server (http://raptorx.uchicago.edu), the DeepMetaPSICOV server (http://bioinf.cs.ucl.ac.uk/psipred/), and the SPOT-Contact server (https://sparks-lab.org/server/spot-contact/). A sequence length limit (500) is imposed by the capacity of coevolution analysis servers, and may be circumvented if installing and running coevolution analysis locally.

These methods use multiple sequence alignments to compute the probability P_{ij} of residue pair (i, j) forming spatial contact. Based on the matrix of predicted P_{ij}, a protein residue contact map can be built with residues as nodes and pairwise contacts as edges weighted by P_{ij}. By default, we do not apply any threshold cutoff to P_{ij} for defining contacts (unless networks with unweighted edges are required by some node centrality algorithms in NetworkX, where we remove edges with $P_{ij}<0.1$, and set weight to 1 for the remaining edges).

3. Network analysis of protein contact map

By treating a protein contact map as a network of nodes and edges, we calculate various node centrality scores to predict key residues as hotspots for disease mutations.

A simple score to measure node centrality is a weighted node degree that accounts for the nearest neighbor interactions (denoted W_1):

$$W_{1,i} = \sum_{k \neq i} P_{ik} \qquad (1)$$

To include indirect couplings beyond the nearest neighbors, we calculate the node degree based on the n'th power of the contact probability matrix (denoted W_n):

$$W_{n,i} = \sum_{k \neq i} P_{ik} W_{n-1,k} = \sum_{k \neq i} P^{n}{}_{ik} \qquad (2)$$

As n goes to infinity, W_n converges to the eigenvector of P matrix with the highest eigenvalue $\lambda_{\max}$ (denoted W_∞):

$$PW_{\infty} = \lambda_{\max} W_{\infty} \quad (3)$$

Among various W_n, W_2 can be interpreted as the node degrees of a new network based on a neighborhood similarity matrix S as follows (denoted W_s):

$$S_{ij} = \sum_{k \neq i,\, j} P_{ik} P_{jk}, \; W_{s,i} = \sum_{k \neq i} S_{ik} \quad (4)$$

In this study we use five network scores (W_1, W_2, W_3, W_∞ and W_s) as predictive features for node importance. Additionally, we exploit 13 network centrality metrics as calculated by the NetworkX package (see Table 1). To allow meaningful comparison of scores between proteins, the scores of each protein are sorted and their ranking percentiles are linearly transformed to values between 0 and 1.

Table 1. Network centrality scores as implemented in the NetworkX package

(see https://networkx.org/documentation/stable/reference/algorithms/centrality.html)

Symbol	Centrality name	Definition
C1	degree_centrality	Corresponding to W_1
C2	eigenvector_centrality	Corresponding to W_∞
C3	closeness_centrality	Closeness centrality of a node u is the reciprocal of the average shortest path distance to u over all n-1 reachable nodes.
C4	betweenness_centrality	Betweenness centrality of a node u is the sum of the fraction of all-pairs shortest paths that pass through u.
C5	current_flow_closeness_centrality	Current-flow closeness centrality is a variant of closeness centrality based on effective resistance between nodes in a network.
C6	current_flow_betweenness_centrality	Current-flow betweenness centrality is based on an electrical current model for information spreading.
C7	communicability_betweenness_centrality	Communicability betweenness centrality is based on the number of walks connecting every pair of nodes.
C8	load_centrality	Load centrality of a node u is the fraction of all shortest paths that pass through u.
C9	subgraph_centrality	Subgraph centrality of a node u is the sum of weighted closed walks of all lengths starting and ending at u.
C10	harmonic_centrality	Harmonic centrality of a node u is the sum of the reciprocal of the shortest path distances from all other nodes to u.

C11	second_order_centrality	Second order centrality of a node u is the standard deviation of the return times to u of a perpetual random walk on G.
C12	laplacian_centrality	Laplacian Centrality of a node u is measured by the drop in the Laplacian Energy after deleting u from the graph.
C13	katz_centrality_numpy	Katz centrality computes the centrality for a node u based on the centrality of its neighbors. It is a generalization of the eigenvector centrality.

4. Sequences-based GNM

For comparison, we implemented Bulter et al's sequence-based GNM [44]. The original structure-based Gaussian network model (GNM) represents a protein structure as an elastically connected network of residues to obtain the equilibrium fluctuations of residues. In the absence of a structure, the sequence-based GNM (Seq-GNM) treats coevolving residue pairs as contacting pairs.

To construct the Kirchhoff matrix (denoted *K*), each non-bonded residue pair is assigned a value of -1 times its contact probability. The bonded residue pairs (i, i+1) are assigned -1 to enforce local chain connectivity. The diagonal elements of *K* are assigned so that the sum of each row and column is zero:

$$K_{ij} = \begin{cases} -P_{ij} & i \neq j \\ \sum_{k \neq i} P_{ik} & i = j \end{cases} \qquad (5)$$

The vibrational thermal fluctuations of residues are evaluated by inverting the Kirchhoff matrix (or summing over the modes as weighted by $1/\lambda_m$). The per-residue mean-square fluctuations (MSF), which are proportional to the crystallographic B factors, are given as follows:

$$MSF_i \propto K_{ii}^{-1} = \sum_{m>0} \frac{V_{mi}^{\ 2}}{\lambda_m} \quad (6)$$

where the eigen-decomposition of K gives eigenvectors V_m and eigenvalues λ_m that satisfy:

$$KV_m = \lambda_m V_m \quad (7)$$

Low-MSF residues correspond to rigid cores or hinges of dynamical importance [44].

As an alternative way to evaluate node importance using GNM, we perform a perturbation-based hotspot analysis as follows: For mode *m*, calculate how much its eigenvalue changes ($\delta\lambda_{m,i}$) in response to a perturbation at a chosen residue position i [23,24,51] (i.e., by uniformly weakening the contacts with residue i). Then compute $\delta\lambda_i = \sum_m \delta\lambda_{m,i}$ to assess the dynamic importance of this residue position [52]. High-$\delta\lambda_i$ residues correspond to sites highly sensitive to local perturbations that mimic mutations.

The above two GNM-based scores are combined with the other network scores for machine learning.

5. ESM based variant prediction

For comparison with our method, we use a deep-learning variant predictor based on a large protein language model (ESM). We downloaded and installed the ESM package and pretrained models from https://github.com/facebookresearch/esm. Since our dataset consists of known variants (from HumVar) and added non-conserved sites (with specific mutations unknown), we simulate the mutational effects on each site by introducing Alanine substitution if

the wildtype residue is not an Alanine and Glycine substitution otherwise. Then we process the mutated sequence with 5 pretrained ESM models (esm1v_t33_650M_UR90S_1, esm1v_t33_650M_UR90S_2, esm1v_t33_650M_UR90S_3, esm1v_t33_650M_UR90S_4, and esm1v_t33_650M_UR90S_5), which predict the difference in the probability of observing the widetype residue and the mutant residue at a given site [48]. We record the predictions of five ESM models as separate features to be optimally integrated via machine learning.

6. AlphaFold for structural prediction

We downloaded predicted structures for the 107 proteins from AlphaFold DB (https://alphafold.ebi.ac.uk/). A residue contact probability matrix is constructed based on the predicted structures as follows:

$$P_{ij} = \frac{1}{1+e^{d_{ij}-10}} \quad (8)$$

where d_{ij} is the distance between residues i and j, and 10 Å is used as a soft cutoff distance. We then use this contact probability matrix to perform the same network analysis as in the sequences-based method and for optimization with machine learning.

7. Foldx for structural refinement and Alanine scanning analysis

FoldX program [53] was downloaded from https://foldxsuite.crg.eu/. We use the RepairPDB command to refine the AlphaFold-predicted models (by fixing bad torsion angles

and Van der Waals clashes). Then we use the AlaScan command to mutate each residue to Ala and calculate the resulting changes in Gibbs free energies which are then used as a feature to predict hotspots of disease mutations.

8. Machine learning algorithms

We use the following machine learning methods of the scikit-learn package (https://scikit-learn.org/stable/) to learn optimal combinations of all features to predict if a given site is deleterious or neutral mutation site:

Random Forest Classifier (RF) (sklearn.ensemble.RandomForestClassifier): A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. We tune the following hyper-parameters: max_depth, n_estimators, max_features.

Gradient Boosting Classifier (GB) (sklearn.ensemble.GradientBoostingClassifier): This algorithm builds an additive model in a forward stage-wise fashion. In each stage a regression tree is fit on the negative gradient of the loss function, e.g. binary log loss. We tune the following hyper-parameters: n_estimators, max_depth, max_features.

Extreme Gradient Boosting Classifier (XGB) (xgboost.XGBClassifier): This algorithm is an optimized distributed version of gradient boosting designed to be highly efficient, flexible and portable. We tune the following hyper-parameters: n_estimators, max_depth, reg_alpha, reg_lambda.

These three methods were chosen because they have performed successfully in machine learning contests in Kaggle (see https://www.packtpub.com/product/the-kaggle-book/9781801817479). They are also relatively cheap to train and optimize compared with the deep learning methods.

We use Optuna (https://optuna.org/) for hyper-parameter tuning of the above algorithms. We have run Optuna multiple times to ensure the resulting best metric is reproducible.

Results and discussion

This study explores how to systematically utilize the coevolution information from multiple sequence alignments to model and analyze a protein as a residue contact network beyond the scope of GNM. To this end, we first use coevolution analysis to construct a protein residue contact map with edges weighted by the predicted contact probability; then we exploit an array of 20 network-based scores to assess the node importance as predictors for disease mutation sites; finally we evaluate the predictive power of these scores individually and collectively (using machine learning) based on a subset of 107 protein sequences and their variants from the HumVar database. For comparison, we also evaluate alternative methods based on predicted protein structures, a physics-based force field, and protein language models.

1. Evaluation of individual network scores

Based on the protein residue contact maps built from three coevolution analysis tools (DeepMetaPSICOV, RaptorX, and SPOT-Contact), we applied network analysis to calculate 20 network scores (see Table 2), measuring node centrality using various different algorithms (see Methods). These scores include simple weighted node degrees for n-hop nearest neighbors (see Methods) and more sophisticated centrality metrics (see Table 1), along with 2 seq-GNM based scores (MSF and $\delta\lambda$, see Methods). We evaluate the performance of each score using the AUC of ROC for the testing set, which provides a balanced evaluation of sensitivity and specificity (see Table 2). More specifically, we sort all testing-set variants by a particular score and predict a

variant deleterious/neutral if its score is above/below a cutoff value. This results in an ROC curve from which we have calculated AUC (see Table 2).

Overall, DeepMetaPSICOV (max AUC=0.80) and SPOT-Contact (max AUC=0.81) perform slightly better than RaptorX (max AUC=0.78). Interestingly, simple weighted node degrees (W_1, W_2, and W_3) perform better than those more complex centrality scores (see Table 2). When computing node degrees, going beyond the nearest neighbors seems to improve the prediction slightly (see Table 2). Two GNM-based scores perform similarly but slightly worse than the weighted node degrees (see Table 2). Among those NetworkX-based scores (see Table 1), C5, C11 and C12 outperform the others, while those betweenness-based scores (C4, C6, and C8) underperform (see Table 2). Therefore, the functional importance of a node/residue pertains more to its role as a highly-connected hub than as an information bottleneck of the shortest paths.

Table 2. Evaluation of 20 network scores based on protein residue contact maps constructed from 3 coevolution analysis tools (DeepMetaPSICOV, RaptorX, and SPOT-Contact) and AlphaFold-predicted structures

Score	AUC* of DeepMetaPSICOV	AUC* of RaptorX	AUC* of SPOT-Contact	AUC* of AlphaFold
C1	0.74	0.76	0.73	0.82
C2	0.73	0.74	0.76	0.77
C3	0.76	0.73	0.69	0.73
C4	0.64	0.54	0.60	0.58
C5	0.78	0.76	0.79	0.80
C6	0.63	0.58	0.67	0.60
C7	0.75	0.61	0.72	0.74
C8	0.64	0.54	0.60	0.58
C9	0.77	0.76	0.74	0.78
C10	0.75	0.73	0.68	0.75
C11	0.79	0.76	0.77	0.80
C12	0.77	0.77	0.79	0.83
C13	0.73	0.73	0.76	0.76
$\delta\lambda$	0.79	0.76	0.78	0.83
MSF	0.79	0.76	0.78	0.80

W_1	0.79	0.77	0.78	**0.83**
W_2	0.80	0.78	0.80	0.83
W_3	**0.80**	**0.78**	**0.81**	0.82
W_∞	0.80	0.74	0.79	0.77
W_s	0.80	0.78	0.80	0.83
FoldX				0.68

* The AUC is calculated based on the ROC for all variants of the 28 testing set proteins. Alternatively, we also calculated AUCs based on the ROCs of individual proteins and their summary statistics (see Table S1).

For comparison with alternative methods, we evaluated the performance of variant prediction by five pre-trained protein language models (ESM, see Methods), and the resulting AUC varies between 0.79 and 0.81, which are comparable to the network scores (see Table 2). For further comparison with structures-based methods, we also performed network analysis based on protein structures as predicted by AlphaFold (see Methods). Overall, the structures-based scores (max AUC=0.83) perform slightly better than the sequences-based scores. This may be partly due to the structure-based contact maps (see Eq. 8) being more sharply defined than the fuzzier contact-probability-based contact maps. Notably, when structural information is used, our network analysis performs significantly better than a physics-based force field (FoldX) with AUC=0.68. Taken together, these findings support the usefulness of individual sequences-based network centrality scores in predicting important residues on par with alternative more sophisticated methods.

To further understand the different accuracies of the above scores, we explore the relationships between them by evaluating the pairwise Pearson correlations (PC) (see Table 3). W_1, W_2, W_3, W_∞, W_s , MSF and $\delta\lambda$ are highly correlated (with PC$\geq$0.93 for DeepMetaPSICOV, PC$\geq$0.84 for SPOT-Contact, PC$\geq$0.86 for AlphaFold), although their correlations are somewhat weaker for RaptorX. Among the NetworkX-based scores (see Table 1), C5, C11 and C12 are

also highly correlated with the above scores. Such strong correlations support the attribution of higher AUC of these scores (see Table 2) to their capturing the same essential features (i.e. high node degrees) of those important nodes. In contrast, the betweenness-based scores (C4, C6, and C8) do not correlate well with the above scores, which is consistent with their lower AUC (see Table 2).

Table 3. Pearson correlations between network scores (row 1, 2, 3 and 4 correspond to results of DeepMetaPSICOV, SPOT-Contact, RaptorX and AlphaFold, respectively)

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	W_1	W_2	W_3	W_∞	W_s	MSF	$\delta\lambda$
C1	1.00	0.57	0.74	0.63	0.72	0.40	0.82	0.63	0.84	0.79	0.88	0.84	0.54	0.87	0.85	0.83	0.80	0.85	0.87	0.87
C2	0.57	1.00	0.43	0.21	0.75	0.18	0.47	0.21	0.66	0.42	0.67	0.76	0.98	0.65	0.72	0.75	0.80	0.71	0.65	0.65
C3	0.74	0.43	1.00	0.64	0.63	0.36	0.72	0.64	0.73	0.96	0.81	0.57	0.43	0.63	0.63	0.63	0.61	0.64	0.64	0.63
C4	0.63	0.21	0.64	1.00	0.34	0.42	0.64	1.00	0.40	0.65	0.53	0.37	0.21	0.46	0.42	0.40	0.38	0.43	0.47	0.46
C5	0.72	0.75	0.63	0.34	1.00	0.46	0.67	0.34	0.74	0.58	**0.91**	**0.86**	0.76	**0.80**	**0.83**	**0.84**	**0.84**	**0.83**	**0.80**	**0.80**
C6	0.40	0.18	0.36	0.42	0.46	1.00	0.63	0.42	0.21	0.29	0.45	0.33	0.20	0.45	0.41	0.38	0.34	0.42	0.46	0.45
C7	0.82	0.47	0.72	0.64	0.67	0.63	1.00	0.64	0.73	0.71	0.80	0.68	0.46	0.76	0.75	0.74	0.71	0.76	0.77	0.76
C8	0.63	0.21	0.64	1.00	0.34	0.42	0.64	1.00	0.40	0.65	0.53	0.37	0.21	0.46	0.42	0.40	0.37	0.43	0.47	0.46
C9	0.84	0.66	0.73	0.40	0.74	0.21	0.73	0.40	1.00	0.79	0.85	0.79	0.63	0.77	0.81	0.82	0.82	0.81	0.78	0.77
C10	0.79	0.42	0.96	0.65	0.58	0.29	0.71	0.65	0.79	1.00	0.79	0.59	0.41	0.65	0.65	0.64	0.63	0.66	0.66	0.65
C11	0.88	0.67	0.81	0.53	0.91	0.45	0.80	0.53	0.85	0.79	1.00	**0.85**	0.68	**0.84**	**0.85**	**0.85**	**0.84**	**0.86**	**0.85**	**0.84**
C12	0.84	0.76	0.57	0.37	0.86	0.33	0.68	0.37	0.79	0.59	**0.85**	1.00	0.74	**0.91**	**0.94**	**0.94**	**0.92**	**0.92**	**0.90**	**0.91**
C13	0.54	0.98	0.43	0.21	0.76	0.20	0.46	0.21	0.63	0.41	0.68	0.74	1.00	0.63	0.69	0.73	0.78	0.69	0.63	0.63
W_1	0.87	0.65	0.63	0.46	0.80	0.45	0.76	0.46	0.77	0.65	0.84	0.91	0.63	**1.00**	**0.98**	**0.97**	**0.93**	**0.98**	**1.00**	**1.00**
W_2	0.85	0.72	0.63	0.42	0.83	0.41	0.75	0.42	0.81	0.65	0.85	0.94	0.69	**0.98**	**1.00**	**1.00**	**0.97**	**1.00**	**0.99**	**0.98**
W_3	0.83	0.75	0.63	0.40	0.84	0.38	0.74	0.40	0.82	0.64	0.85	0.94	0.73	**0.97**	**1.00**	**1.00**	**0.99**	**0.99**	**0.97**	**0.97**
W_∞	0.80	0.80	0.61	0.38	0.84	0.34	0.71	0.37	0.82	0.63	0.84	0.92	0.78	**0.93**	**0.97**	**0.99**	**1.00**	**0.97**	**0.93**	**0.93**
W_s	0.85	0.71	0.64	0.43	0.83	0.42	0.76	0.43	0.81	0.66	0.86	0.92	0.69	**0.98**	**1.00**	**0.99**	**0.97**	**1.00**	**0.99**	**0.98**
MSF	0.87	0.65	0.64	0.47	0.80	0.46	0.77	0.47	0.78	0.66	0.85	0.90	0.63	**1.00**	**0.99**	**0.97**	**0.93**	**0.99**	**1.00**	**1.00**
C1	1.00	0.58	0.68	0.55	0.73	0.53	0.85	0.55	0.86	0.73	0.90	0.74	0.57	0.80	0.78	0.77	0.67	0.80	0.82	0.80
C2	0.58	1.00	0.44	0.20	0.88	0.40	0.51	0.20	0.61	0.38	0.74	0.82	0.99	0.70	0.76	0.79	0.88	0.74	0.68	0.70
C3	0.68	0.44	1.00	0.61	0.55	0.38	0.69	0.61	0.73	0.96	0.78	0.42	0.45	0.50	0.52	0.53	0.52	0.55	0.56	0.50
C4	0.55	0.20	0.61	1.00	0.29	0.46	0.63	1.00	0.37	0.62	0.49	0.26	0.20	0.37	0.35	0.34	0.30	0.37	0.40	0.37
C5	0.73	0.88	0.55	0.29	1.00	0.57	0.67	0.29	0.71	0.49	**0.88**	**0.91**	**0.87**	**0.84**	**0.88**	**0.89**	**0.87**	**0.87**	**0.82**	**0.84**
C6	0.53	0.40	0.38	0.46	0.57	1.00	0.73	0.46	0.33	0.31	0.54	0.55	0.40	0.64	0.61	0.59	0.49	0.62	0.66	0.64
C7	0.85	0.51	0.69	0.63	0.67	0.73	1.00	0.63	0.75	0.68	0.82	0.66	0.51	0.76	0.75	0.73	0.64	0.77	0.80	0.76
C8	0.55	0.20	0.61	1.00	0.29	0.46	0.63	1.00	0.37	0.62	0.49	0.26	0.20	0.37	0.35	0.34	0.30	0.37	0.40	0.37
C9	0.86	0.61	0.73	0.37	0.71	0.33	0.75	0.37	1.00	0.77	0.87	0.66	0.59	0.68	0.71	0.72	0.68	0.73	0.71	0.68
C10	0.73	0.38	0.96	0.62	0.49	0.31	0.68	0.62	0.77	1.00	0.76	0.38	0.38	0.47	0.49	0.49	0.47	0.52	0.54	0.47
C11	0.90	0.74	0.78	0.49	0.88	0.54	0.82	0.49	0.87	0.76	1.00	**0.79**	0.73	**0.80**	**0.82**	**0.83**	**0.78**	**0.83**	**0.82**	**0.80**
C12	0.74	0.82	0.42	0.26	0.91	0.55	0.66	0.26	0.66	0.38	**0.79**	1.00	0.80	**0.91**	**0.93**	**0.92**	**0.83**	**0.90**	**0.86**	**0.91**
C13	0.57	0.99	0.45	0.20	0.87	0.40	0.51	0.20	0.59	0.38	0.73	0.80	1.00	0.69	0.75	0.78	0.88	0.73	0.67	0.69
W_1	0.80	0.70	0.50	0.37	0.84	0.64	0.76	0.37	0.68	0.47	0.80	0.91	0.69	**1.00**	**0.98**	**0.97**	**0.84**	**0.98**	**0.98**	**1.00**
W_2	0.78	0.76	0.52	0.35	0.88	0.61	0.75	0.35	0.71	0.49	0.82	0.93	0.75	**0.98**	**1.00**	**1.00**	**0.90**	**1.00**	**0.97**	**0.98**
W_3	0.77	0.79	0.53	0.34	0.89	0.59	0.73	0.34	0.72	0.49	0.83	0.92	0.78	**0.97**	**1.00**	**1.00**	**0.92**	**0.99**	**0.96**	**0.97**
W_∞	0.67	0.88	0.52	0.30	0.87	0.49	0.64	0.30	0.68	0.47	0.78	0.83	0.88	**0.84**	**0.90**	**0.92**	**1.00**	**0.90**	**0.86**	**0.84**

W_s	0.80 0.74 0.55 0.37 0.87 0.62 0.77 0.37 0.73 0.52 0.83 0.90 0.73 **0.98 1.00 0.99 0.90 1.00 0.98 0.98**
MSF	0.82 0.68 0.56 0.40 0.82 0.66 0.80 0.40 0.71 0.54 0.82 0.86 0.67 **0.98 0.97 0.96 0.86 0.98 1.00 0.98**
C1	1.00 0.63 0.58 0.22 0.64 0.22 0.29 0.22 0.75 0.65 0.71 0.78 0.55 0.88 0.87 0.85 0.58 0.87 0.68 0.73
C2	0.63 1.00 0.64 0.15 0.82 0.23 0.24 0.15 0.66 0.66 0.78 0.70 0.93 0.60 0.65 0.67 0.82 0.64 0.70 0.55
C3	0.58 0.64 1.00 0.44 0.79 0.35 0.39 0.44 0.64 0.93 0.88 0.51 0.68 0.48 0.50 0.50 0.64 0.50 0.69 0.47
C4	0.22 0.15 0.44 1.00 0.32 0.70 0.67 1.00 0.11 0.38 0.39 0.06 0.22 0.18 0.14 0.14 0.30 0.15 0.37 0.17
C5	0.64 0.82 0.79 0.32 1.00 0.45 0.43 0.32 0.65 0.72 **0.96** 0.71 **0.86** 0.63 0.65 0.66 0.79 0.65 **0.81** 0.62
C6	0.22 0.23 0.35 0.70 0.45 1.00 0.70 0.70 0.12 0.25 0.45 0.17 0.29 0.32 0.28 0.26 0.40 0.28 0.51 0.31
C7	0.29 0.24 0.39 0.67 0.43 0.70 1.00 0.67 0.34 0.34 0.48 0.28 0.29 0.30 0.28 0.27 0.40 0.29 0.49 0.39
C8	0.22 0.15 0.44 1.00 0.32 0.70 0.67 1.00 0.11 0.38 0.39 0.06 0.22 0.18 0.14 0.14 0.30 0.15 0.37 0.17
C9	0.75 0.66 0.64 0.11 0.65 0.12 0.34 0.11 1.00 0.71 0.72 0.71 0.61 0.66 0.69 0.69 0.62 0.69 0.66 0.64
C10	0.65 0.66 0.93 0.38 0.72 0.25 0.34 0.38 0.71 1.00 0.83 0.55 0.67 0.52 0.53 0.53 0.66 0.54 0.72 0.51
C11	0.71 0.78 0.88 0.39 0.96 0.45 0.48 0.39 0.72 0.83 1.00 0.68 **0.82** 0.64 0.65 0.65 0.77 0.65 **0.83** 0.63
C12	0.78 0.70 0.51 0.06 0.71 0.17 0.28 0.06 0.71 0.55 0.68 1.00 0.64 0.76 0.77 0.77 0.63 0.77 0.67 0.70
C13	0.55 0.93 0.68 0.22 0.86 0.29 0.29 0.22 0.61 0.67 0.82 0.64 1.00 0.53 0.58 0.61 0.85 0.58 0.72 0.52
W_1	0.88 0.60 0.48 0.18 0.63 0.32 0.30 0.18 0.66 0.52 0.64 0.76 0.53 **1.00 0.98 0.97** 0.69 **0.98 0.81 0.85**
W_2	0.87 0.65 0.50 0.14 0.65 0.28 0.28 0.14 0.69 0.53 0.65 0.77 0.58 **0.98 1.00 0.99** 0.73 **1.00 0.80 0.83**
W_3	0.85 0.67 0.50 0.14 0.66 0.26 0.27 0.14 0.69 0.53 0.65 0.77 0.61 **0.97 0.99 1.00** 0.75 **0.99 0.80 0.82**
W_∞	0.58 0.82 0.64 0.30 0.79 0.40 0.40 0.30 0.62 0.66 0.77 0.63 0.85 0.69 0.73 0.75 1.00 0.73 **0.88** 0.67
W_s	0.87 0.64 0.50 0.15 0.65 0.28 0.29 0.15 0.69 0.54 0.65 0.77 0.58 **0.98 1.00 0.99** 0.73 1.00 **0.81 0.83**
MSF	0.68 0.70 0.69 0.37 0.81 0.51 0.49 0.37 0.66 0.72 0.83 0.67 0.72 **0.81 0.80 0.80 0.88 0.81** 1.00 0.79
C1	1.00 0.87 0.77 0.38 0.95 0.42 0.82 0.38 0.90 0.81 0.96 0.98 0.85 0.97 0.97 0.96 0.86 0.97 0.95 0.97
C2	0.87 1.00 0.78 0.26 0.91 0.32 0.81 0.26 0.98 0.79 0.90 0.89 0.99 0.87 0.91 0.93 0.99 0.91 0.91 0.87
C3	0.77 0.78 1.00 0.52 0.79 0.37 0.74 0.52 0.82 0.97 0.83 0.72 0.78 0.71 0.74 0.75 0.78 0.74 0.79 0.71
C4	0.38 0.26 0.52 1.00 0.33 0.61 0.55 1.00 0.30 0.51 0.39 0.30 0.27 0.30 0.30 0.29 0.27 0.30 0.33 0.30
C5	0.95 0.91 0.79 0.33 1.00 0.46 0.85 0.33 0.92 0.79 **0.99 0.96** 0.91 **0.95 0.96 0.96 0.91 0.96 1.00 0.95**
C6	0.42 0.32 0.37 0.61 0.46 1.00 0.67 0.61 0.33 0.31 0.46 0.40 0.33 0.41 0.39 0.37 0.33 0.39 0.46 0.41
C7	0.82 0.81 0.74 0.55 0.85 0.67 1.00 0.55 0.83 0.72 0.85 0.81 0.81 0.80 0.82 0.83 0.81 0.82 0.84 0.80
C8	0.38 0.26 0.52 1.00 0.33 0.61 0.55 1.00 0.30 0.51 0.39 0.30 0.27 0.30 0.30 0.29 0.27 0.30 0.33 0.30
C9	0.90 0.98 0.82 0.30 0.92 0.33 0.83 0.30 1.00 0.84 0.93 0.91 0.96 0.89 0.93 0.95 0.97 0.93 0.92 0.89
C10	0.81 0.79 0.97 0.51 0.79 0.31 0.72 0.51 0.84 1.00 0.84 0.75 0.79 0.73 0.77 0.78 0.79 0.77 0.80 0.73
C11	0.96 0.90 0.83 0.39 0.99 0.46 0.85 0.39 0.93 0.84 1.00 **0.95** 0.90 **0.94 0.95 0.95 0.90 0.95 0.99 0.94**
C12	0.98 0.89 0.72 0.30 0.96 0.40 0.81 0.30 0.91 0.75 **0.95** 1.00 0.88 **1.00 1.00 0.99 0.89 0.99 0.96 1.00**
C13	0.85 0.99 0.78 0.27 0.91 0.33 0.81 0.27 0.96 0.79 0.90 0.88 1.00 0.85 0.90 0.92 0.99 0.90 0.91 0.85
W_1	0.97 0.87 0.71 0.30 0.95 0.41 0.80 0.30 0.89 0.73 0.94 1.00 0.85 **1.00 0.99 0.97 0.86 0.98 0.95 1.00**
W_2	0.97 0.91 0.74 0.30 0.96 0.39 0.82 0.30 0.93 0.77 0.95 1.00 0.90 **0.99 1.00 1.00 0.91 1.00 0.96 0.99**
W_3	0.96 0.93 0.75 0.29 0.96 0.37 0.83 0.29 0.95 0.78 0.95 0.99 0.92 **0.97 1.00 1.00 0.93 1.00 0.96 0.97**
W_∞	0.86 0.99 0.78 0.27 0.91 0.33 0.81 0.27 0.97 0.79 0.90 0.89 0.99 **0.86 0.91 0.93 1.00 0.91 0.91 0.86**
W_s	0.97 0.91 0.74 0.30 0.96 0.39 0.82 0.30 0.93 0.77 0.95 0.99 0.90 **0.98 1.00 1.00 0.91 1.00 0.96 0.98**
MSF	0.95 0.91 0.79 0.33 1.00 0.46 0.84 0.33 0.92 0.80 0.99 0.96 0.91 **0.95 0.96 0.96 0.91 0.96 1.00 0.95**

In summary, by evaluating 20 network scores individually, we have found a wide range of performance with AUC varying from 0.54 to 0.81 (see Table 2). The top-performing scores seem to correlate strongly with each other, so they must have captured a common aspect of node centrality that is relevant to functional importance (e.g. high local connectivity instead of high

betweenness). Interestingly, the two GNM-based scores, despite measuring distinct dynamic properties (MSF measures thermal fluctuations while $\delta\lambda$ measures sensitivity to local perturbations), are also strongly correlated with each other and those degree-based network scores. Therefore, to speed up the variant prediction workflow we only need to compute those simpler weighted node degrees as features without significantly losing accuracy.

2. Combining all network scores to predict variant hotspots by machine learning

To optimize the predictive power of the above network-based scores based on three coevolution analysis methods (or AlphaFold), we have employed machine learning algorithms (see Methods) to take them as input features, train a binary classifier which predicts if a residue position is linked to neutral or deleterious variants (using first 79 proteins as training set), and then test its prediction using the remaining 28 proteins as testing set. We use the AUC of ROC as the metric for assessing the prediction quality of the trained classifier.

To evaluate the protein residue contact maps constructed by each method, we combine all network scores based on the contact maps predicted by the same method (see Table 2) for machine learning. The resulting AUC of each coevolution analysis method (DeepMetaPSICOV, RaptorX, and SPOT-Contact) is 0.81, 0.80, and 0.82, respectively (see Table 4), which are slightly better than the best AUC of individual scores (0.78~0.81, see Table 2). The lack of substantial improvement may be due to high correlations among the scores (see Table 3) which could reduce the effectiveness of ensemble learning. For comparison, we also trained and tested classifiers using the AlphaFold-predicted contact maps, and alternative classifiers based on

protein language models (see Methods). Both alternative methods give comparable yet slightly better AUC (0.83). Similar to our finding, Butler et al reported AUC of 0.81 after combining the B-factors of Seq-GNM with evolutionary features [44].

Table 4. **Evaluation of classifiers trained by 3 machine learning algorithms (RF, GB and XGB, see Methods) based on the protein residue contact maps constructed from 3 coevolution analysis tools (DeepMetaPSICOV, RaptorX, and SPOT-Contact), AlphaFold-predicted structures, and protein language models (ESM).**

Sources of input features	AUC of RF	AUC of GB	AUC of XGB
DeepMetaPSICOV	0.81	0.81	0.81
RaptorX	0.80	0.80	0.80
SPOT-Contact	0.82	0.82	0.82
AlphaFold	0.83	0.83	0.83
ESM	0.83	0.83	0.83
All 3 coevolution methods	0.84	0.84	0.84
All 3 coevolution methods (w/o C1-C13)	0.82	0.82	0.83
All 3 coevolution methods and ESM	0.89	0.89	0.89
AlphaFold and ESM	0.88	0.88	0.88

To further boost the prediction performance, we have sought to combine the network scores of all three coevolution analysis methods for machine learning, resulting in better AUC (0.84) which slightly outperform both AlphaFold and ESM (0.83). To assess the added value of including 13 NetworkX-based centrality scores (see Table 1), we have performed an ablation study that excludes them in machine learning, and found slightly lower AUC (0.82~0.83). So it is possible to speed up the calculation without significantly reducing accuracy. Taken together, our findings support the power of combining an array of different network scores from different coevolution analysis tools to optimize the prediction in the framework of ensemble learning.

To further explore how well our method complements alternative methods, we have combined all the network scores with the ESM scores in machine learning. Encouragingly, we have obtained markedly improved AUC (0.89), which is comparable to machine learning that combines the AlphaFold-based network scores with the ESM scores (AUC=0.88).

For comparison with other studies, Butler et al showed that Seq-GNM combined with evolutionary parameters attained a sensitivity of 0.84 and a specificity of 0.66 [44]. PolyPhen-2 achieved a sensitivity of 0.73 and a specificity of 0.8 on the HumVar datasets [47]. While using different training and testing datasets, we have attained competitive results with a sensitivity of 0.82 and a specificity of 0.80 (using all the network scores from three coevolution analysis tools and the ESM scores). For more direct comparison, we also evaluated PolyPhen-2 based on the same 28 testing-set proteins and their variants, and obtained an AUC of 0.85, which is close to our method (see Table 4). However, this metric is likely positively biased since PolyPhen-2 has been trained on the HumVar dataset.

In summary, via extensive machine learning, we have demonstrated the power of using an ensemble of sequences-based network scores calculated by different co-evolution analysis tools to accurately predict deleterious mutation sites. Although some network scores are highly correlated (see Table 3) and they vary widely in accuracy (see Table 2), these scores seem to be sufficiently diverse to allow effective ensemble learning when combined.

3. Case studies:

To illustrate the biomedical significance of our predictions of variant sites with network scores, we discuss in details the following four proteins from our dataset.

Glucose-6-phosphate exchanger (Uniprot id: O43826): As an inorganic phosphate and glucose-6-phosphate antiporter, it transports cytoplasmic glucose-6-phosphate into the lumen of the endoplasmic reticulum and translocates inorganic phosphate in the opposite direction. Being involved in glucose production through glycogenolysis and gluconeogenesis, it plays a central role in homeostatic regulation of blood glucose levels. It is linked to diseases like congenital disorder of glycosylation and glycogen storage disease (see https://www.uniprot.org/uniprotkb/O43826/entry#function).

The AlphaFold-predicted structure forms a dimer of transmembrane helical domains with most deleterious mutation sites concentrated inside the central core while those non-conserved residues (i.e. neutral mutation sites) are mostly located on the periphery (see Fig 1c). The contact maps predicted by three coevolution analysis tools all agree well with the contact map based on the AlphaFold structure (see Fig 1a) (except that RaptorX omitted many local contacts in residues 1-200). As a result, the network centrality scores (W_3) also agree well between these methods (see Fig 1b), although the coevolution-based network scores are generally noisier (with more spikes) than the structure-based scores (see Fig 1b). Different network scores calculated from the same contact map are also highly similar (see Fig 1d) despite being based on different algorithms. For example, scores of $\delta\lambda$ and MSF agree very well (see Fig 1d). Encouragingly, those residues identified with high network scores are primarily within the central core (inside each domain or in the inter-domain hinge region), thus overlapping with most deleterious mutations (see Fig 1c). Among those top-10% hotspot residues (see Fig 1c), mutations Y24H, N27K, R28H, G88D, G149E, P153L, and G339C were implicated in causing glycogen storage disease [54] . Two of these mutations (R28H and G149E) were found to exhibit undetectable

microsomal glucose-6-phosphate transport activity in transient expression studies[55], thus confirming their functional importance.

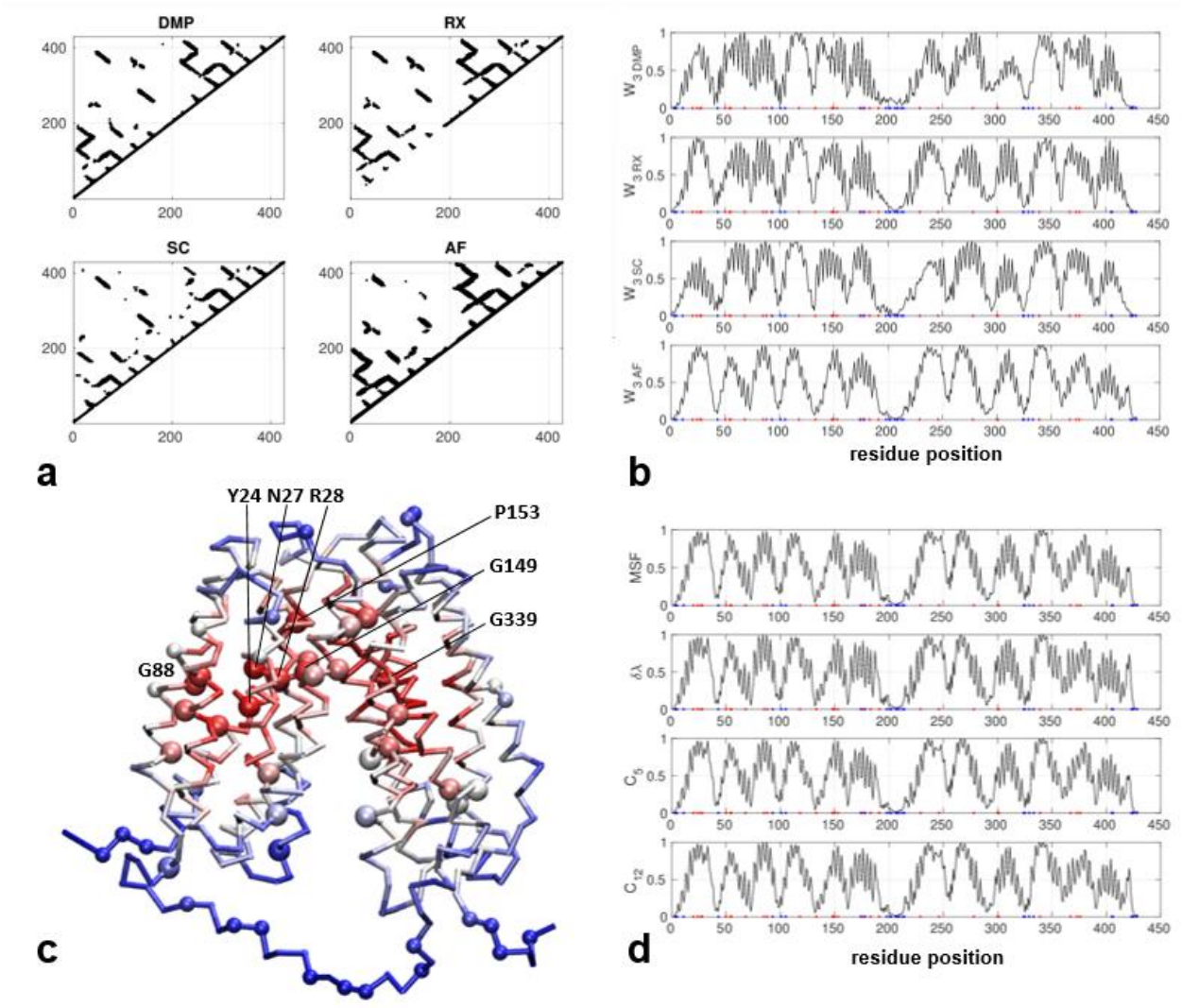


Figure 1. **Results for Glucose-6-phosphate exchanger (Uniprot id: O43826)**: (a) Four contact maps constructed from coevolution analysis by DeepMetaPSICOV (DMP), RaptorX (RX), SPOT-Contact (SC), and the predicted structure by AlphaFold (AF) (only those contacts with probability >0.1 are shown). (b) W_3 scores for all residue positions based on the contact maps in (a), where red and blue dots mark residues with deleterious and neutral mutations, respectively. (c) Predicted structure by AlphaFold as colored by W_3 scores (red/blue for high/low values),

where residues with deleterious and neutral mutations are shown as large and small balls, respectively (G20, Y24, N27, R28, G50, S54, S55, G68, L85, G88, W118, Q133, A148, G149, G150, P153, C176, C183, P191, L229, W246, I278, R300, H301, G339, A367, A373, G376, see https://www.uniprot.org/uniprotkb/O43826/variant-viewer). (d) Four other network scores (MSF, $\delta\lambda$, C5 and C12) for all residue positions based on the contact maps in (a).

Presenilin-1 (Uniprot id: P49768): As the catalytic subunit of the gamma-secretase complex, it catalyzes the intramembrane cleavage of integral membrane proteins such as Notch receptors. It is involved in various diseases including a familial early-onset form of Alzheimer disease and a form of frontotemporal dementia (see https://www.uniprot.org/uniprotkb/P49768/entry#function).

The AlphaFold-predicted structure consists of two closely packed helical domains with most deleterious mutations clustered inside the core domain while the non-conserved residues are mostly located on the N-terminal loop (residues 1-70) and the inter-domain linker (residues 300-370) (see Fig 2c). The active site [56] (D257 and D385) is also located in the core domain (colored green in Fig 2c). The contact maps predicted by three coevolution analysis methods all resemble the contact map based on the predicted structure (see Fig 2a) (except that RaptorX omitted local contacts in residues 1-100). As a result, the network scores agree well between them in the helical domains (see Fig 2c), but with more differences in the flexible regions (residues 1-70 and 300-370). Reassuringly, those residues identified by high network scores are primarily clustered within the central core overlapping with most deleterious mutations, while the flexible N-terminal and linker feature low scores consistent with low sequence conservation (see Fig 2c). Among those top 10% hotspot residues (see Fig 2c), mutations at C92, V96, A231,

M233, L235, A246, L250, S390, L392, and C410 were found to cause loss of function and altered amyloid-beta production [57] : C92S led to loss of protease function and increased Abeta42 levels. V96F caused loss of protease activity. A231T/V and M233T led to decreased protease activity, altered amyloid-beta production and increased amyloid-beta 42/amyloid-beta 40 ratio. L235P/R and S390I abolished protease activity. A246E and L250S abolished protease activity and increased amyloid-beta 42/amyloid-beta 40 ratio. L392V resulted in reduced proteolysis, altered amyloid-beta production and increased amyloid-beta 42/amyloid-beta 40 ratio. C410I reduced proteolysis. Since most of these residues are not near the active site, their effects on protease activity are likely allosteric.

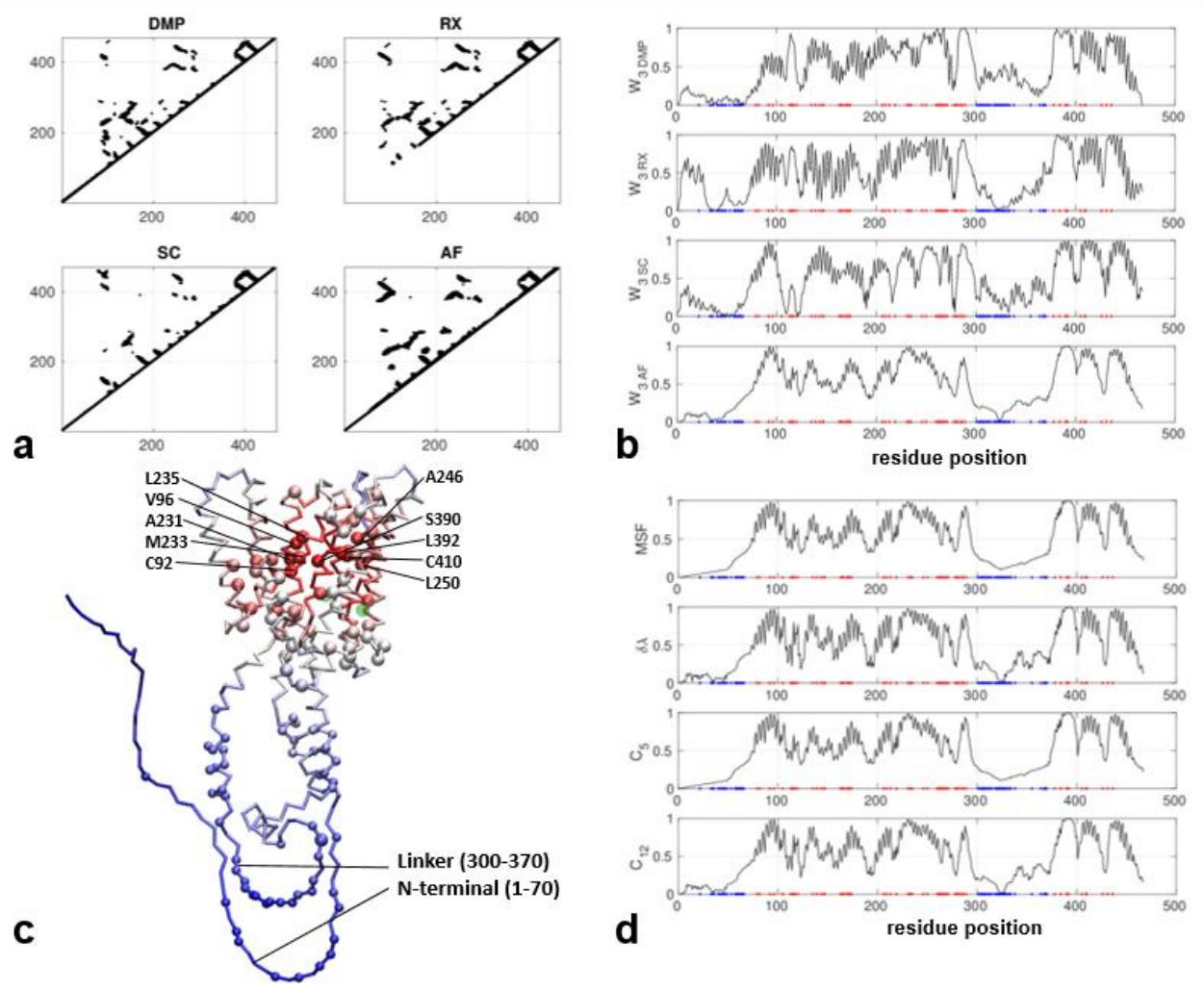


Figure 2. **Results for Presenilin-1 (Uniprot id: P49768)**: (a) Four contact maps constructed from coevolution analysis by DeepMetaPSICOV (DMP), RaptorX (RX), SPOT-Contact (SC), and the predicted structure by AlphaFold (AF) (only those contacts with probability >0.1 are shown). (b) W_3 scores for all residue positions based on the contact maps in (a), where red and blue dots mark residues with deleterious and neutral mutations, respectively. (c) Predicted structure by AlphaFold as colored by W_3 scores (red/blue for high/low values), where residues with deleterious and neutral mutations are shown as large and small balls, respectively (A79, V82, C92, V96, F105, L113, Y115, T116, P117, E120, N135, M139, I143, M146, T147, H163, W165, L166, S169, L171, L173, L174, G206, G209, I213, L219, A231, M233, L235, A246, L250, A260, L262, C263, P264, G266, P267, R269, L271, R278, E280, L282, A285, L286, S289, D333, G378, G384, S390, L392, N405, A409, C410, A426, A431, P436, see https://www.uniprot.org/uniprotkb/ P49768/variant-viewer), and active-site residues are colored in green. (d) Four other network scores (MSF, δλ, C5 and C12) for all residue positions based on the contact maps in (a).

b(0,+)-type amino acid transporter 1 (Uniprot id: P82251): It forms a functional transporter complex that mediates the electrogenic exchange between cationic amino acids and neutral amino acids. Its dysfunction is linked to Cystinuria, an autosomal disorder characterized by impaired epithelial cell transport of cystine and dibasic amino acids in the proximal renal tubule and gastrointestinal tract (see https://www.uniprot.org/uniprotkb/P82251/entry#function).

The AlphaFold-predicted structure consists of a helical domain with deleterious mutations concentrating inside the core domain while those non-conserved residues are mostly located on the domain periphery (N-terminal and C-terminal helices) (see Fig 3c). The active site

consists of residues 43-47 and 233 and is also located in the core domain (colored green in Fig 3c). The contact maps predicted by three coevolution analysis tools are all similar to the contact map based on the AlphaFold structure (see Fig 3a) (except that RaptorX omitted some local contacts in residues 1-200). As a result, the network scores agree well between these methods (see Fig 3b). Reassuringly, those residues identified with high network scores are primarily within the central core and overlap with most deleterious mutations, while the peripheral regions feature low scores consistent with low sequence conservation. Among those top-10% hotspot residues (see Fig 3c), mutations I44T, A126T, and W230R were implicated in Cystinuria. *In vitro* measurements showed W230R has almost no transport activity, and it was proposed that W230 serves as a gate between two substrate-binding pockets and undergoes conformational changes to enable amino acid transport [58] . Although the A126T mutation is mildly dysfunctional [59], it is notable among a cluster of conserved residues with small sidechains in the contact regions of transmembrane helices, hinting for its possible role in helix-helix association and relative motions.

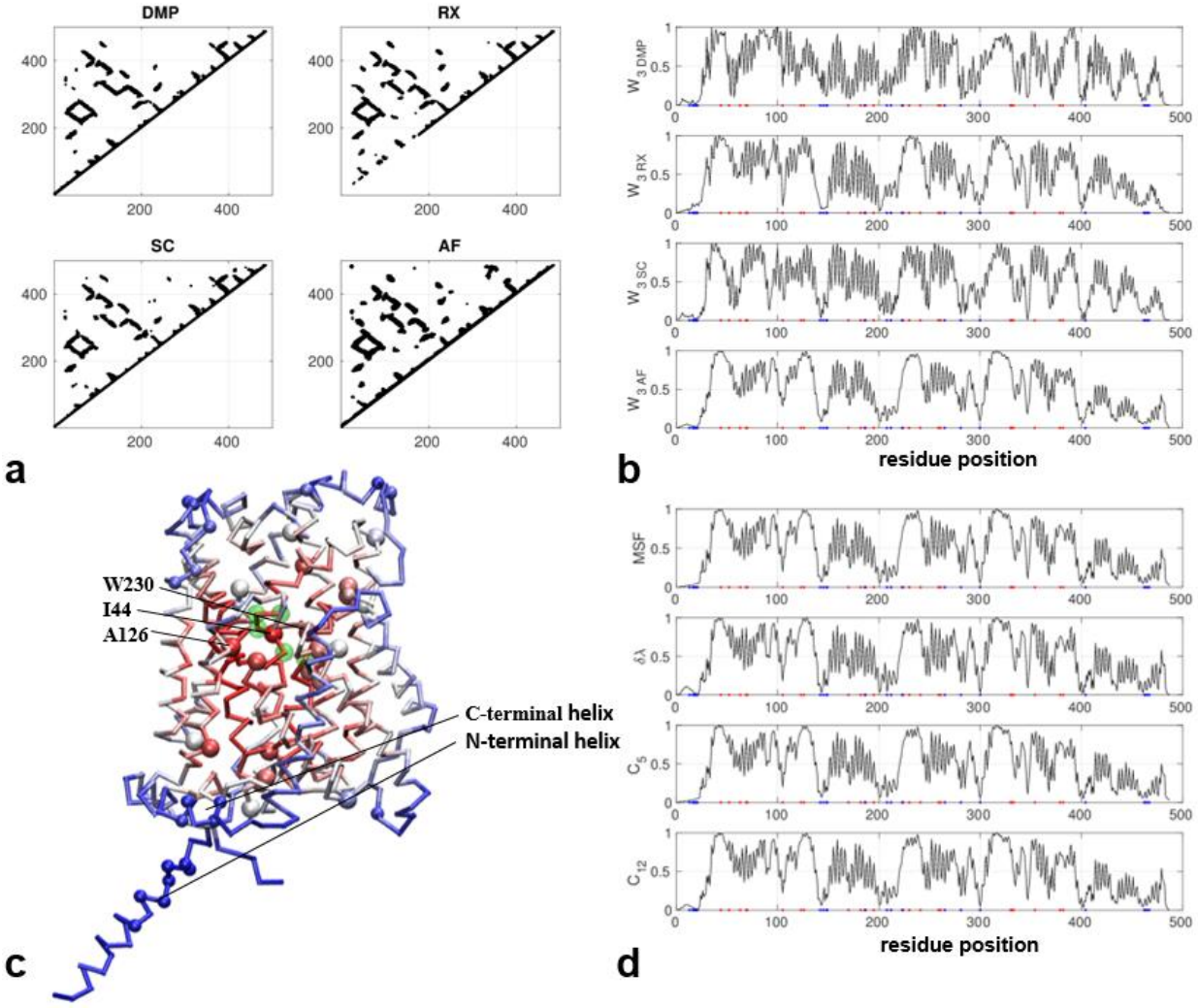


Figure 3. **Results for amino acid transporter 1 (Uniprot id: P82251)**: (a) Four contact maps constructed from coevolution analysis by DeepMetaPSICOV (DMP), RaptorX (RX), SPOT-Contact (SC), and the predicted structure by AlphaFold (AF) (only those contacts with probability >0.1 are shown). (b) W_3 scores for all residue positions based on the contact maps in (a), where red and blue dots mark residues with deleterious and neutral mutations, respectively. (c) Predicted structure by AlphaFold as colored by W_3 scores (red/blue for high/low values), where residues with deleterious and neutral mutations are shown as large and small balls, respectively (V142,L223,I44,P52,G63,W69,A70,G105,T123,A126,V170,A182,I187,G195, A224,W230,I241,G259,P261,V330,A331,R333,A354,S379,A382, see

https://www.uniprot.org/uniprotkb/ P82251/variant-viewer), and active-site residues are colored in green. (d) Four other network scores (MSF, δλ, C5 and C12) for all residue positions based on the contact maps in (a).

Lipoprotein lipase (Uniprot: P06858): As a key enzyme in triglyceride metabolism, it catalyzes the hydrolysis of triglycerides from circulating chylomicrons and very low density lipoproteins, thus playing an important role in lipid clearance from the blood stream, lipid utilization and storage (see https://www.uniprot.org/uniprotkb/P06858/entry#function).

The AlphaFold-predicted structure consists of an N-terminal helix, a central α/β domains, and a C-terminal β domain. Most deleterious mutations are concentrated inside the central domain while the non-conserved residues are mostly located on the periphery (including N-terminal helix and C-terminal domain) (see Fig 4c). The active site is comprised of a catalytic triad of S159, D183, and H268 [60] in the central domain (colored green in Fig 4c). The contact maps predicted by three coevolution analysis methods are similar to the contact map based on the AlphaFold structure (see Fig 4a). As a result, the network scores agree well between these methods (see Fig 4b) with minor differences in peripheral regions (such as the N-terminal helix). As predicted, those residues identified with high network scores are primarily within the central domain overlapping with most deleterious mutations, while the peripheral N-terminal helix and C-terminal domain feature low scores consistent with low sequence conservation (see Fig 4c). Notably, some of them are found at the interface between the central domain and the C-terminal domain (circled in Fig 4c), possibly mediating inter-domain motions. Among those top-10% hotspot residues (see Fig 4c), T128, G132, H163, G169, G181, D183, P184, A185, D207, V208, H210, G222, V227, D231, I232, P234 and S271 are known to harbor pathogenic mutations in

Hyperlipoproteinemia 1, an autosomal recessive metabolic disorder characterized by defective breakdown of dietary fats. Both H163 and G169 lie in helix 4 that constitutes part of the highly conserved beta-epsilon serine-alpha folding motif which is near S159 of the active site. Supporting their functional relevance, mutations H163R and G169E were found to abolish the enzymatic activity [61] . Near D183 (one of the catalytic triad), mutations G181S and P184R were found to abolish the catalytic activity [62] . Further from D183, conserved substations D207E and H210Q abolished the enzyme activity [63], and mutations D231E, I232S and P234L led to loss of the catalytic function [64] . These mutations may disrupt allosteric interactions with the central catalytic domain. Another conservative mutation S271T (near D183) also led to loss of enzyme activity [65]. Taken together, these residues may function by directly or indirectly coupling to the active site.

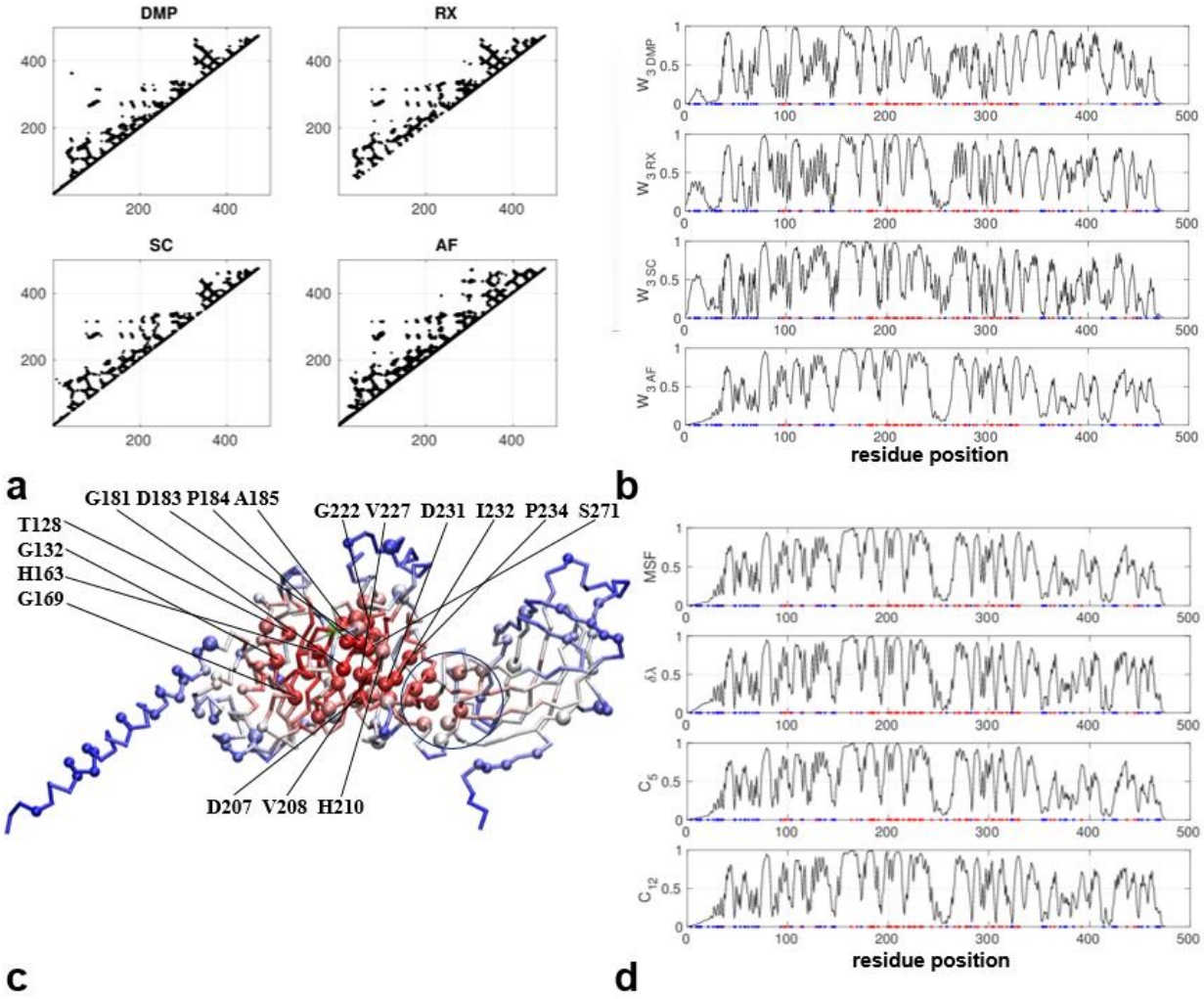


Figure 4. Results for Lipoprotein lipase (Uniprot id: P06858): (a) Four contact maps constructed from coevolution analysis by DeepMetaPSICOV (DMP), RaptorX (RX), SPOT-Contact (SC), and the predicted structure by AlphaFold (AF) (only those contacts with probability >0.1 are shown). (b) W_3 scores for all residue positions based on the contact maps in (a), where red and blue dots mark residues with deleterious and neutral mutations, respectively. (c) Predicted structure by AlphaFold as colored by W_3 scores (red/blue for high/low values), where residues with deleterious and neutral mutations are shown as large and small balls, respectively

(H71,A427,D36,N70,V96,A98,R102,W113,T128,G132,H163,G169,G181,D183,P184,A185,

G186,E190,S199,D201,A203,D207,V208,H210,G215, S220,I221,G222, K225,V227,D231, I232,P234,C243,I252,C266,R270,S271,D277,S278,L279,S286,Y289,F297,L303,C305, C310,L313,N318,S325,M328,L330,A361,S365,L392,E437,E437,C445,E448, see https://www.uniprot.org/uniprotkb/P06858/variant-viewer), and active-site residues are colored in green. (d) Four other network scores (MSF, δλ, C5 and C12) for all residue positions based on the contact maps in (a).

Conclusion

To conclude, we have combined machine learning, network analysis, and protein language models to develop a sequences-based variant site prediction method based on the protein residue contact networks which incorporate sequential, structural, dynamic, and interaction information simultaneously:

1. We build protein residue networks by exploiting three different state-of-the-art coevolution analysis tools (RaptorX, DeepMetaPSICOV, and SPOT-Contact) that complement each other.

2. We use three powerful machine learning algorithms (Random Forest, Gradient Boosting, and Extreme Gradient Boosting) to optimally combine 20 network centrality scores to accurately predict key residues as hot spots for disease mutations.

3. We train and validate our method using a dataset of 107 proteins rich in disease mutations, demonstrating its high accuracy in distinguishing between deleterious and neutral sites (with AUC of ROC ~ 0.84). Further improvement can be achieved after combining our method with the ESM-based method.

This study has established a useful strategy of combining an ensemble of network scores based on different coevolution analysis methods via machine learning to predict key variants sites of relevance to disease mutations. The code and dataset are made available to public to enable future developments and applications (see https://simtk.org/projects/hotspots).

For future work, it will be interesting to go beyond contact map predictions by integrating other scores derived from the co-evolution analysis (for example, see refs [66-68]) in our workflow, which may further boost the accuracy of variant site prediction.

References

1. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, et al. (2021) Highly accurate protein structure prediction with AlphaFold. Nature 596: 583-589.
2. Baek M, DiMaio F, Anishchenko I, Dauparas J, Ovchinnikov S, et al. (2021) Accurate prediction of protein structures and interactions using a three-track neural network. Science 373: 871-876.
3. Terwilliger TC, Liebschner D, Croll TI, Williams CJ, McCoy AJ, et al. (2023) AlphaFold predictions are valuable hypotheses and accelerate but do not replace experimental structure determination. Nat Methods.
4. Al-Janabi A (2022) Has DeepMind's AlphaFold solved the protein folding problem? Biotechniques 72: 73-76.
5. Medina E, D RL, Sanabria H (2021) Unraveling protein's structural dynamics: from configurational dynamics to ensemble switching guides functional mesoscale assemblies. Curr Opin Struct Biol 66: 129-138.
6. Pak MA, Markhieva KA, Novikova MS, Petrov DS, Vorobyev IS, et al. (2023) Using AlphaFold to predict the impact of single mutations on protein stability and function. PLoS One 18: e0282689.
7. Steinegger M, Soding J (2018) Clustering huge protein sequence sets in linear time. Nat Commun 9: 2542.
8. Burley SK, Berman HM, Bhikadiya C, Bi C, Chen L, et al. (2019) RCSB Protein Data Bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. Nucleic Acids Res 47: D464-D474.
9. Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, et al. (2022) AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Res 50: D439-D444.
10. Bepler T, Berger B (2021) Learning the protein language: Evolution, structure, and function. Cell Syst 12: 654-669 e653.
11. Cheng J, Novati G, Pan J, Bycroft C, Zemgulyte A, et al. (2023) Accurate proteome-wide missense variant effect prediction with AlphaMissense. Science 381: eadg7492.
12. Hassan MS, Shaalan AA, Dessouky MI, Abdelnaiem AE, ElHefnawi M (2019) A review study: Computational techniques for expecting the impact of non-synonymous single nucleotide variants in human diseases. Gene 680: 20-33.
13. Niroula A, Urolagin S, Vihinen M (2015) PON-P2: prediction method for fast and reliable identification of harmful variants. PLoS One 10: e0117380.
14. Pejaver V, Urresti J, Lugo-Martinez J, Pagel KA, Lin GN, et al. (2020) Inferring the molecular and phenotypic impact of amino acid variants with MutPred2. Nat Commun 11: 5918.
15. Qi H, Zhang H, Zhao Y, Chen C, Long JJ, et al. (2021) MVP predicts the pathogenicity of missense variants by deep learning. Nat Commun 12: 510.
16. Yue P, Melamud E, Moult J (2006) SNPs3D: candidate gene and SNP selection for association studies. BMC Bioinformatics 7: 166.
17. Tang H, Thomas PD (2016) Tools for Predicting the Functional Impact of Nonsynonymous Genetic Variation. Genetics 203: 635-647.
18. Katsonis P, Koire A, Wilson SJ, Hsu TK, Lua RC, et al. (2014) Single nucleotide variations: biological impact and theoretical interpretation. Protein Sci 23: 1650-1666.

19. Singh A, Olowoyeye A, Baenziger PH, Dantzer J, Kann MG, et al. (2008) MutDB: update on development of tools for the biochemical analysis of genetic variation. Nucleic Acids Res 36: D815-819.
20. Bromberg Y, Rost B (2007) SNAP: predict effect of non-synonymous polymorphisms on function. Nucleic Acids Res 35: 3823-3835.
21. Ng PC, Henikoff S (2003) SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res 31: 3812-3814.
22. Pers TH, Timshel P, Hirschhorn JN (2015) SNPsnap: a Web-based tool for identification and annotation of matched SNPs. Bioinformatics 31: 418-420.
23. Zheng W, Tekpinar M (2009) Large-scale evaluation of dynamically important residues in proteins predicted by the perturbation analysis of a coarse-grained elastic model. BMC Struct Biol 9: 45.
24. Zheng W, Brooks BR, Doniach S, Thirumalai D (2005) Network of dynamically important residues in the open/closed transition in polymerases is strongly conserved. Structure 13: 565-577.
25. Ponzoni L, Bahar I (2018) Structural dynamics is a determinant of the functional significance of missense variants. Proc Natl Acad Sci U S A 115: 4164-4169.
26. Butler BM, Gerek ZN, Kumar S, Ozkan SB (2015) Conformational dynamics of nonsynonymous variants at protein interfaces reveals disease association. Proteins 83: 428-435.
27. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, et al. (2021) Applying and improving AlphaFold at CASP14. Proteins 89: 1711-1721.
28. Marks DS, Hopf TA, Sander C (2012) Protein structure prediction from sequence variation. Nat Biotechnol 30: 1072-1080.
29. Morcos F, Pagnani A, Lunt B, Bertolino A, Marks DS, et al. (2011) Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proc Natl Acad Sci U S A 108: E1293-1301.
30. Hopf TA, Scharfe CP, Rodrigues JP, Green AG, Kohlbacher O, et al. (2014) Sequence co-evolution gives 3D contacts and structures of protein complexes. Elife 3.
31. Marks DS, Colwell LJ, Sheridan R, Hopf TA, Pagnani A, et al. (2011) Protein 3D structure computed from evolutionary sequence variation. PLoS One 6: e28766.
32. Burger L, van Nimwegen E (2010) Disentangling direct from indirect co-evolution of residues in protein alignments. PLoS Comput Biol 6: e1000633.
33. Jones DT, Buchan DW, Cozzetto D, Pontil M (2012) PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. Bioinformatics 28: 184-190.
34. Halabi N, Rivoire O, Leibler S, Ranganathan R (2009) Protein sectors: evolutionary units of three-dimensional structure. Cell 138: 774-786.
35. Wang S, Sun S, Li Z, Zhang R, Xu J (2017) Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model. PLoS Comput Biol 13: e1005324.
36. Ma J, Wang S, Wang Z, Xu J (2015) Protein contact prediction by integrating joint evolutionary coupling analysis and supervised learning. Bioinformatics 31: 3506-3513.
37. Kandathil SM, Greener JG, Jones DT (2019) Prediction of interresidue contacts with DeepMetaPSICOV in CASP13. Proteins 87: 1092-1099.
38. Jones DT, Singh T, Kosciolek T, Tetchner S (2015) MetaPSICOV: combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins. Bioinformatics 31: 999-1006.
39. Jones DT, Kandathil SM (2018) High precision in protein contact prediction using fully convolutional neural networks and minimal sequence features. Bioinformatics 34: 3308-3315.

685 40. Hanson J, Paliwal K, Litfin T, Yang Y, Zhou Y (2018) Accurate prediction of protein contact maps by
686 coupling residual two-dimensional bidirectional long short-term memory with convolutional
687 neural networks. Bioinformatics 34: 4039-4045.
688 41. Yan W, Yu C, Chen J, Zhou J, Shen B (2020) ANCA: A Web Server for Amino Acid Networks
689 Construction and Analysis. Front Mol Biosci 7: 582702.
690 42. Amitai G, Shemesh A, Sitbon E, Shklar M, Netanely D, et al. (2004) Network analysis of protein
691 structures identifies functional residues. J Mol Biol 344: 1135-1146.
692 43. Velickovic P (2023) Everything is connected: Graph neural networks. Curr Opin Struct Biol 79:
693 102538.
694 44. Butler BM, Kazan IC, Kumar A, Ozkan SB (2018) Coevolving residues inform protein dynamics profiles
695 and disease susceptibility of nSNVs. PLoS Comput Biol 14: e1006626.
696 45. Chasman D, Adams RM (2001) Predicting the functional consequences of non-synonymous single
697 nucleotide polymorphisms: structure-based assessment of amino acid variation. J Mol Biol 307:
698 683-706.
699 46. Gerek ZN, Ozkan SB (2011) Change in allosteric network affects binding affinities of PDZ domains:
700 analysis through perturbation response scanning. PLoS Comput Biol 7: e1002154.
701 47. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, et al. (2010) A method and server
702 for predicting damaging missense mutations. Nat Methods 7: 248-249.
703 48. Meier J, Rao R, Verkuil R, Liu J, Sercu T, et al. (2021) Language models enable zero-shot prediction of
704 the effects of mutations on protein function. bioRxiv: 2021.2007.2009.450648.
705 49. Vihinen M (2020) Problems in variation interpretation guidelines and in their implementation in
706 computational tools. Mol Genet Genomic Med 8: e1206.
707 50. Ashkenazy H, Erez E, Martz E, Pupko T, Ben-Tal N (2010) ConSurf 2010: calculating evolutionary
708 conservation in sequence and structure of proteins and nucleic acids. Nucleic Acids Res 38:
709 W529-533.
710 51. Zheng W, Brooks BR, Thirumalai D (2006) Low-frequency normal modes that describe allosteric
711 transitions in biological nanomachines are robust to sequence variations. Proc Natl Acad Sci U S
712 A 103: 7664-7669.
713 52. Zheng W (2016) Probing the structural dynamics of the SNARE recycling machine based on coarse-
714 grained modeling. Proteins.
715 53. Delgado J, Radusky LG, Cianferoni D, Serrano L (2019) FoldX 5.0: working with RNA, small molecules
716 and a new graphical interface. Bioinformatics 35: 4168-4169.
717 54. Veiga-da-Cunha M, Gerin I, Chen YT, de Barsy T, de Lonlay P, et al. (1998) A gene on chromosome
718 11q23 coding for a putative glucose- 6-phosphate translocase is mutated in glycogen-storage
719 disease types Ib and Ic. Am J Hum Genet 63: 976-983.
720 55. Hiraiwa H, Pan CJ, Lin B, Moses SW, Chou JY (1999) Inactivation of the glucose 6-phosphate
721 transporter causes glycogen storage disease type 1b. J Biol Chem 274: 5532-5536.
722 56. Wolfe MS, Xia W, Ostaszewski BL, Diehl TS, Kimberly WT, et al. (1999) Two transmembrane
723 aspartates in presenilin-1 required for presenilin endoproteolysis and gamma-secretase activity.
724 Nature 398: 513-517.
725 57. Sun L, Zhou R, Yang G, Shi Y (2017) Analysis of 138 pathogenic mutations in presenilin-1 on the in
726 vitro production of Abeta42 and Abeta40 peptides by gamma-secretase. Proc Natl Acad Sci U S A
727 114: E476-E485.
728 58. Yan R, Li Y, Shi Y, Zhou J, Lei J, et al. (2020) Cryo-EM structure of the human heteromeric amino acid
729 transporter b(0,+)AT-rBAT. Sci Adv 6: eaay6379.
730 59. Font MA, Feliubadalo L, Estivill X, Nunes V, Golomb E, et al. (2001) Functional analysis of mutations
731 in SLC7A9, and genotype-phenotype correlation in non-Type I cystinuria. Hum Mol Genet 10:
732 305-316.

60. Emmerich J, Beg OU, Peterson J, Previato L, Brunzell JD, et al. (1992) Human lipoprotein lipase. Analysis of the catalytic triad by site-directed mutagenesis of Ser-132, Asp-156, and His-241. J Biol Chem 267: 4161-4165.
61. Reina M, Brunzell JD, Deeb SS (1992) Molecular basis of familial chylomicronemia: mutations in the lipoprotein lipase and apolipoprotein C-II genes. J Lipid Res 33: 1823-1832.
62. Bruin T, Tuzgol S, van Diermen DE, Hoogerbrugge-van der Linden N, Brunzell JD, et al. (1993) Recurrent pancreatitis and chylomicronemia in an extended Dutch kindred is caused by a Gly154-->Ser substitution in lipoprotein lipase. J Lipid Res 34: 2109-2119.
63. Haubenwallner S, Horl G, Shachter NS, Presta E, Fried SK, et al. (1993) A novel missense mutation in the gene for lipoprotein lipase resulting in a highly conservative amino acid substitution (Asp180-->Glu) causes familial chylomicronemia (type I hyperlipoproteinemia). Genomics 18: 392-396.
64. Gotoda T, Yamada N, Kawamura M, Kozaki K, Mori N, et al. (1991) Heterogeneous mutations in the human lipoprotein lipase gene in patients with familial lipoprotein lipase deficiency. J Clin Invest 88: 1856-1864.
65. Hata A, Emi M, Luc G, Basdevant A, Gambert P, et al. (1990) Compound heterozygote for lipoprotein lipase deficiency: Ser----Thr244 and transition in 3' splice site of intron 2 (AG----AA) in the lipoprotein lipase gene. Am J Hum Genet 47: 721-726.
66. Bisardi M, Rodriguez-Rivas J, Zamponi F, Weigt M (2022) Modeling Sequence-Space Exploration and Emergence of Epistatic Signals in Protein Evolution. Mol Biol Evol 39.
67. Cocco S, Feinauer C, Figliuzzi M, Monasson R, Weigt M (2018) Inverse statistical physics of protein sequences: a key issues review. Rep Prog Phys 81: 032601.
68. Rodriguez-Rivas J, Croce G, Muscat M, Weigt M (2022) Epistatic models predict mutable sites in SARS-CoV-2 proteins and epitopes. Proc Natl Acad Sci U S A 119.

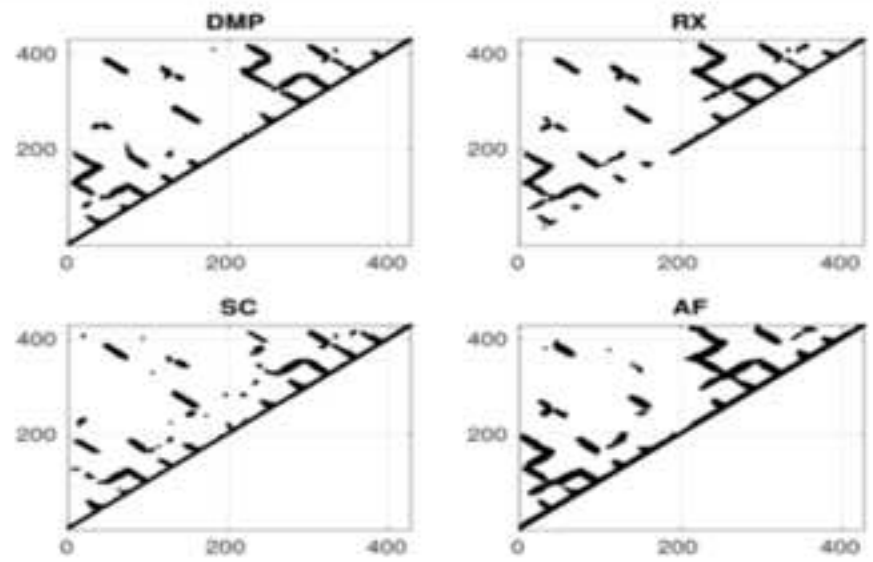
Supporting Information

S1 Table. Evaluation of 20 network scores based on protein residue contact maps constructed from 3 coevolution analysis tools (DeepMetaPSICOV, RaptorX, and SPOT-Contact)

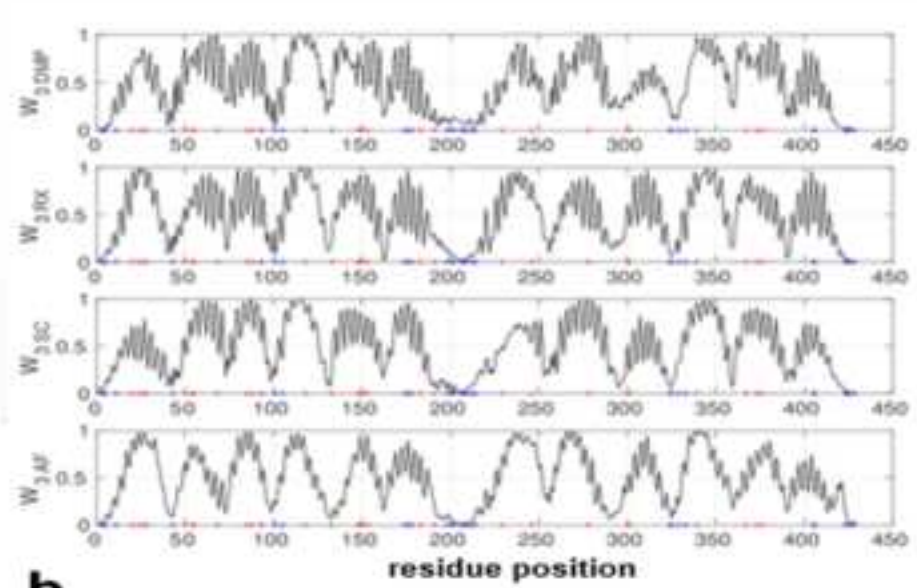
Score	AUC* of DeepMetaPSICOV	AUC* of RaptorX	AUC* of SPOT-Contact
C1	0.75±0.13	0.78±0.14	0.75±0.15
C2	0.75±0.17	0.75±0.19	0.76±0.19
C3	0.78±0.10	0.75±0.15	0.70±0.15
C4	0.65±0.12	0.52±0.15	0.61±0.07
C5	0.78±0.17	0.78±0.17	0.78±0.18
C6	0.63±0.16	0.56±0.19	0.65±0.17
C7	0.77±0.11	0.61±0.20	0.72±0.16
C8	0.65±0.12	0.52±0.15	0.61±0.07
C9	0.79±0.13	0.78±0.16	0.75±0.16
C10	0.76±0.12	0.75±0.14	0.69±0.15
C11	0.80±0.13	0.78±0.17	0.78±0.17
C12	0.78±0.16	0.80±0.14	0.80±0.17
C13	0.75±0.17	0.74±0.18	0.76±0.19
$\delta\lambda$	0.80±0.13	0.78±0.14	0.78±0.16
MSF	0.81±0.14	0.79±0.15	0.80±0.17
W_1	0.81±0.14	0.79±0.15	0.80±0.17
W_2	0.81±0.15	0.77±0.15	0.79±0.17
W_3	0.81±0.14	0.78±0.15	0.80±0.16
W_∞	0.80±0.13	0.77±0.14	0.78±0.16
W_s	0.80±0.13	0.78±0.13	0.78±0.16

* mean ± standard-deviation

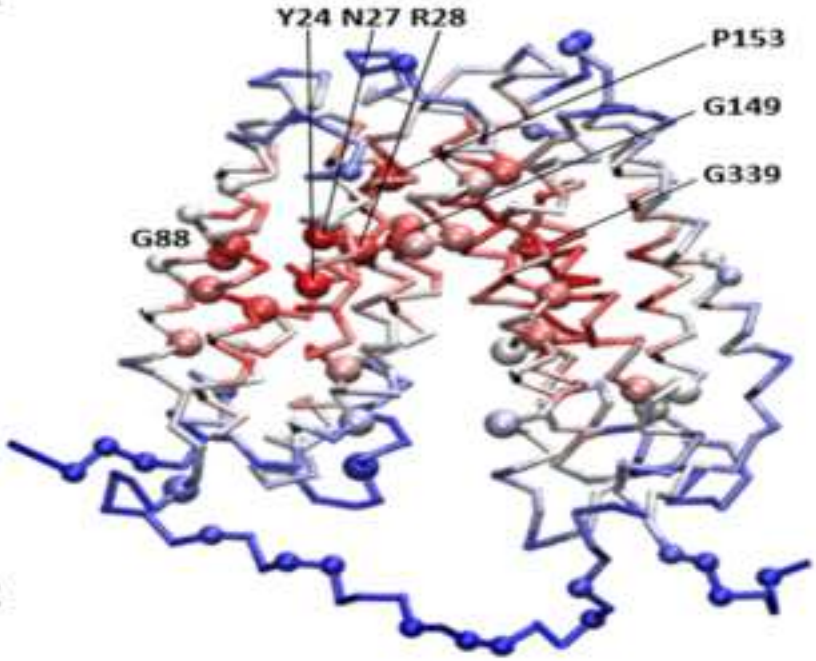
Figure 1



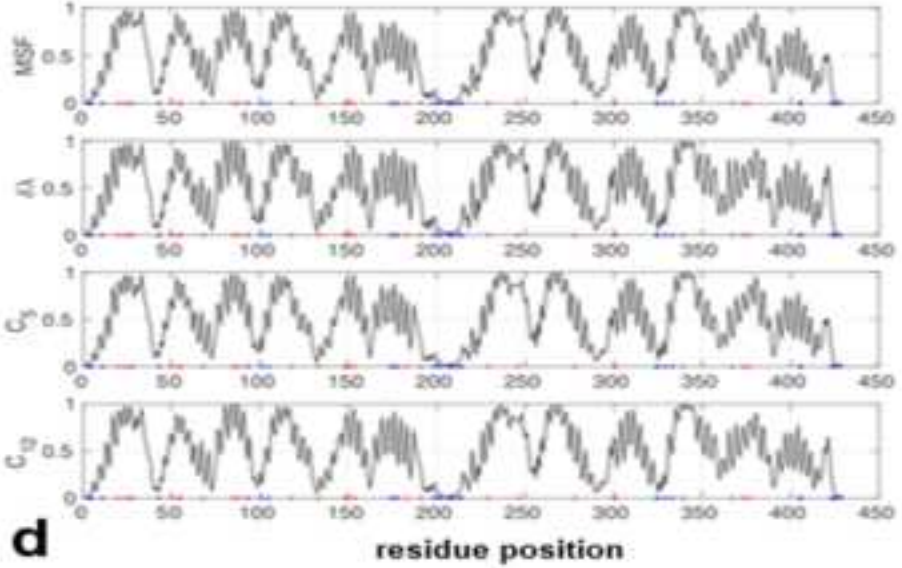
a



b



c



d

Figure 2

Click here to access/download;Figure;Fig2.tif

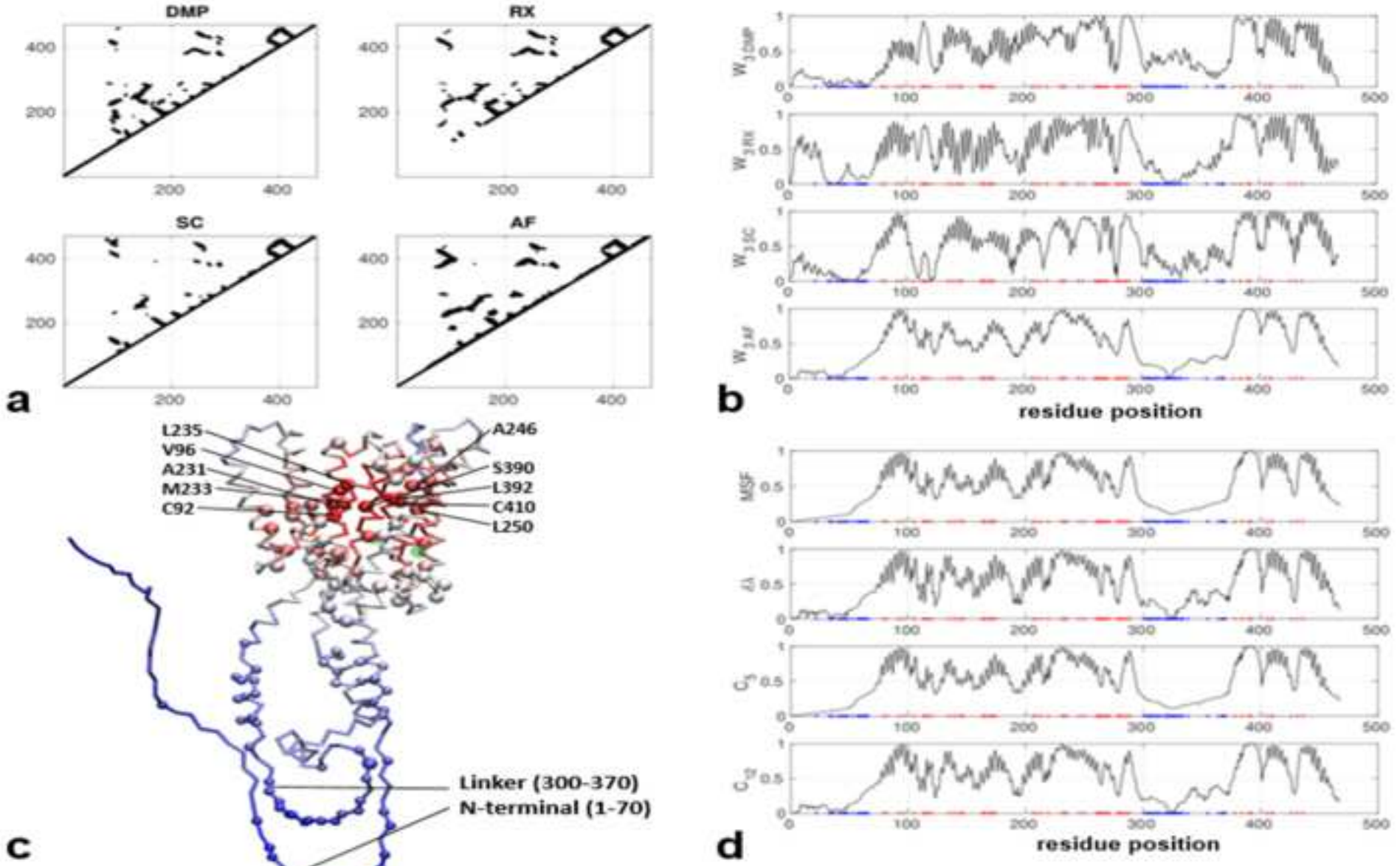
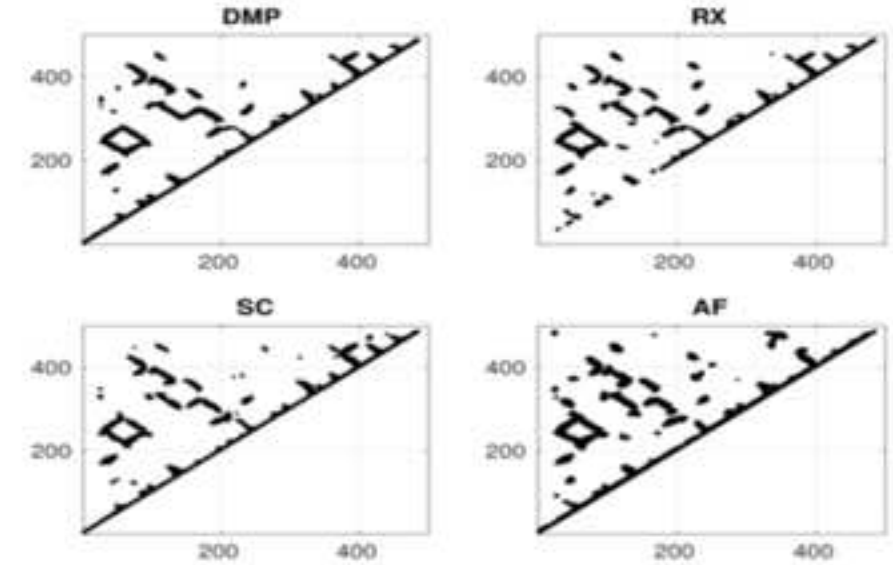
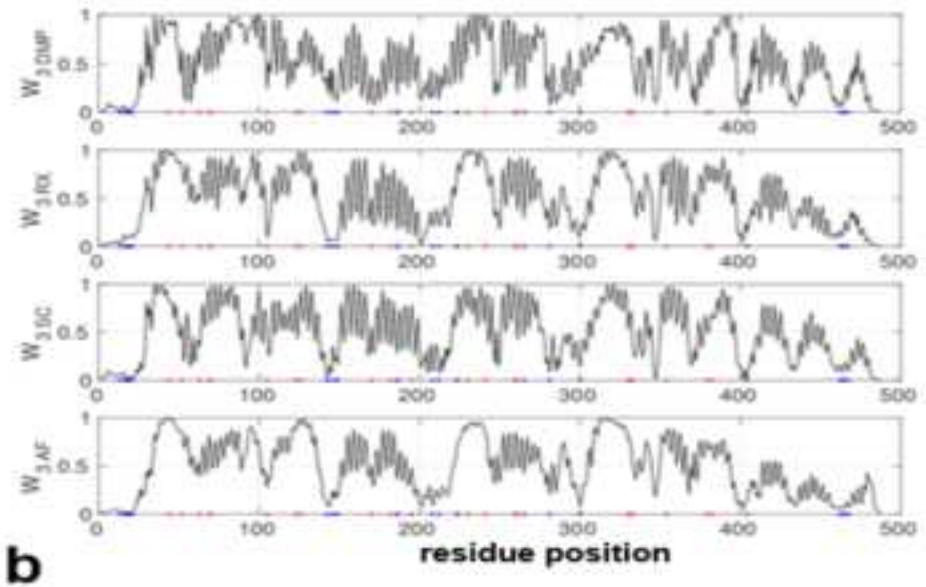


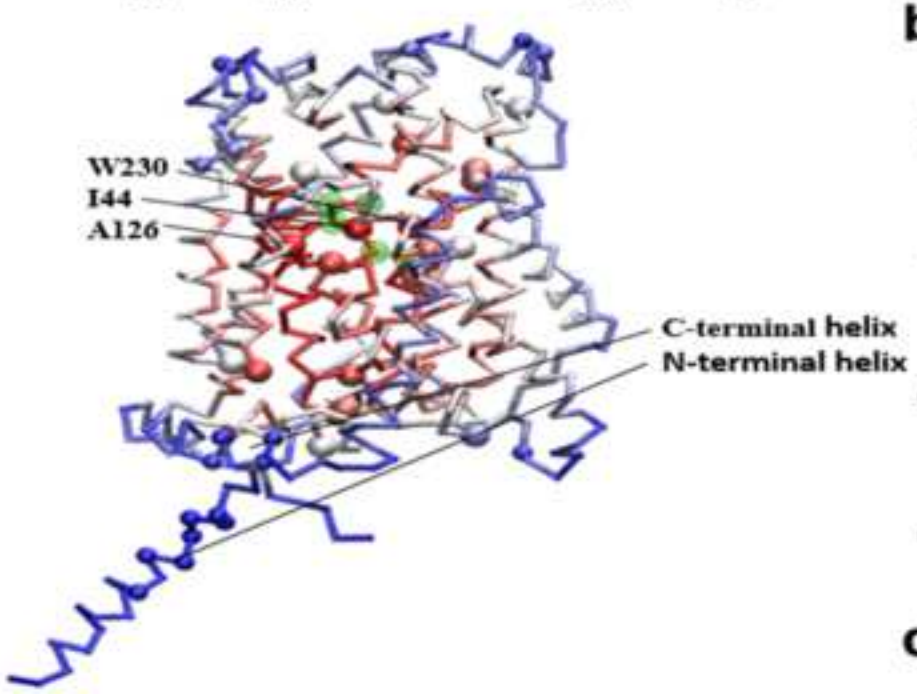
Figure 3



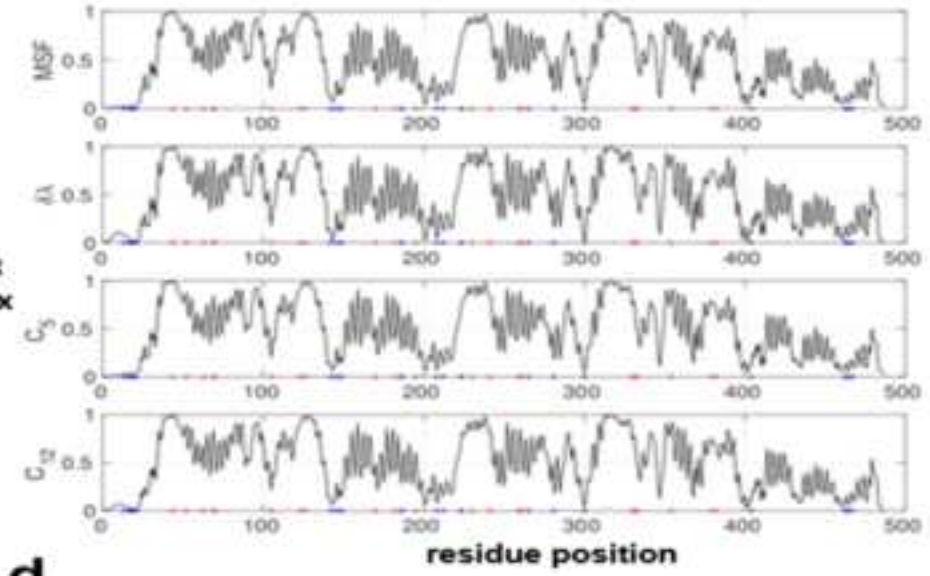
a



b

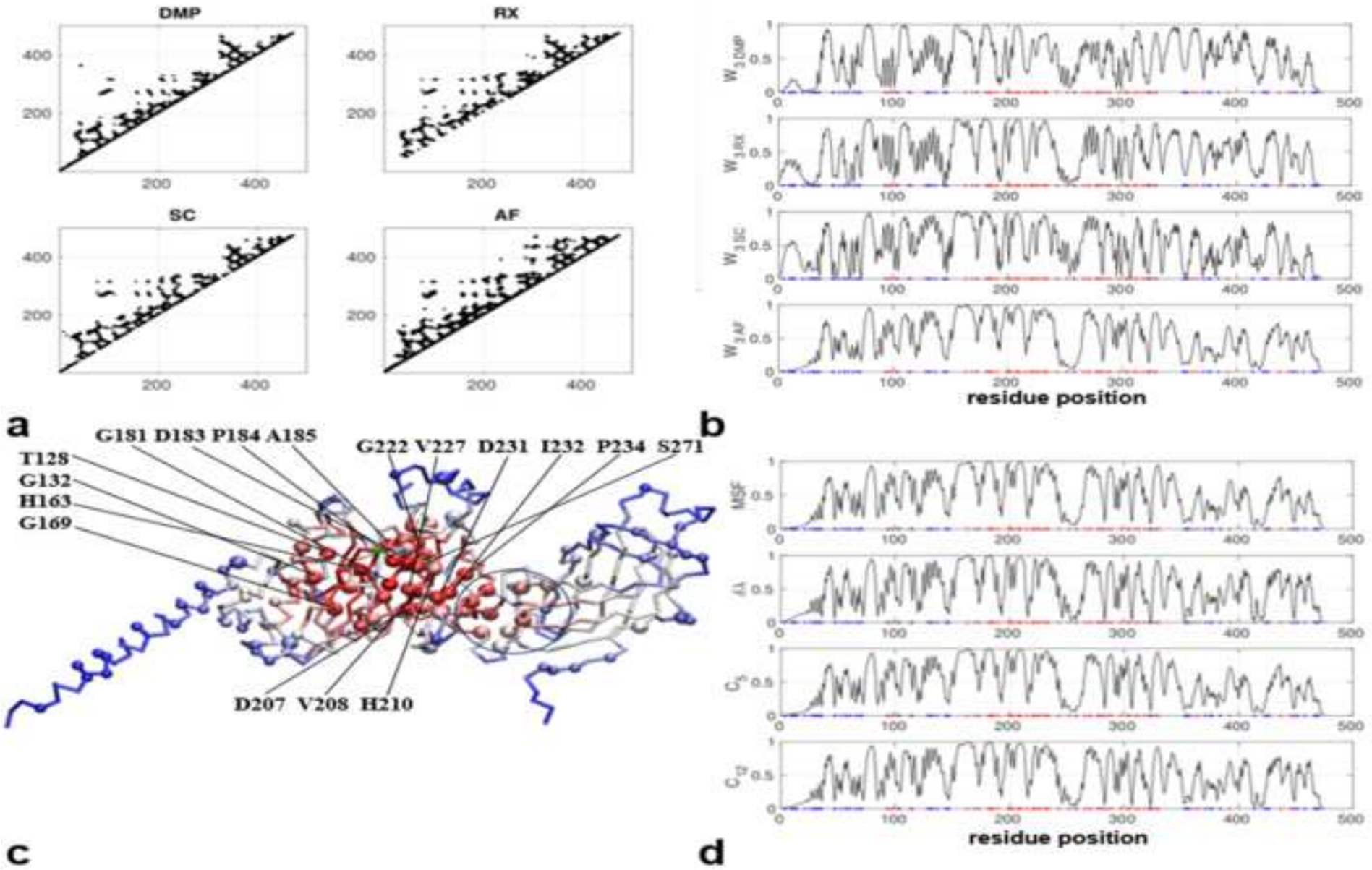


c



d

Figure 4



Predicting hotspots for disease-causing single nucleotide variants using sequences-based coevolution, network analysis, and machine learning

Short title: Predicting nSNVs hotspots with sequences-based machine learning

Wenjun Zheng [1*]

[1] Department of Physics, State University of New York at Buffalo, NY 14260, United States of America

* Corresponding author

E-mail: wjzheng@buffalo.edu (WZ)

Keywords: Centrality, Coevolution, Disease mutations, Machine learning, Protein residue contact network, ~~Protein language model,~~ Single nucleotide variant

Formatted: Font: Not Bold

Abstract

To enable personalized ~~genetics and~~ medicine, it is important yet highly challenging to accurately predict disease-causing mutations in target proteins at high throughput. Previous computational methods have been developed using evolutionary ~~phylogeny~~ information in combination with various biochemical and structural ~~properties~~ features of ~~amino acids~~protein residues to discriminate neutral vs. deleterious mutations. However, the power of these methods is often limited because they either assume known protein structures or ~~do not fully incorporate~~treat residues independently~~, structural, dynamic, and interaction information critical for protein functions. To address these limitations~~without fully considering their global interactions. ~~,~~ To address the above limitations, we build upon recent progress in machine learning, network analysis, and protein language models, and develop a sequences-based variant site prediction workflow based on the protein residue contact networks: 1. We employ and integrate various methods of building protein residue networks using state-of-the-art coevolution analysis tools (~~e.g.,~~ RaptorX, DeepMetaPSICOV, and SPOT-Contact) powered by deep learning. 2. We use machine learning algorithms (~~e.g.,~~ Random Forest, Gradient Boosting, and Extreme Gradient Boosting) to optimally combine ~~13~~ 20 network centrality scores ~~(calculated by NetworkX) with 7 other network scores calculated from the contact probability matrices~~ to jointly predict key residues as hot spots for disease mutations. 3. Using a dataset of 107 proteins rich in disease mutations, we rigorously evaluate the network scores individually and collectively (via machine learning) ~~in comparison with alternative structures-based network scores (using predicted structures by AlphaFold)~~. ~~By optimally combing three coevolution analysis methods and the resulting network scores by machine learning, we are able to discriminate deleterious and neutral mutation sites accurately (AUC of ROC ~ 0.84). Furthermore, by combining our method~~

Formatted: Font color: Auto

Formatted: Font color: Auto

Formatted: Font color: Auto

Formatted: Font color: Auto

Formatted: Font color: Auto

Formatted: Font color: Auto

Formatted: Font color: Auto

Formatted: Font color: Auto

~~with a state-of-the-art predictor of the functional effects of sequence variations based on large protein language models, we have significantly improved the prediction of disease variant sites (AUC ~ 0.89).~~ This work supports a promising strategy of combining an ensemble of network scores based on different coevolution analysis methods (and optionally predictive scores from other methods) via machine learning to predict candidate sites of disease mutations, which will inform downstream applications of disease diagnosis and targeted drug design.

Introduction

The holy grail of structural biology is to solve high-resolution biomolecular structures at the genomic scale to inform mechanistic studies of their functions. Thanks to recent revolutions in computational structural biology (~~e.g.~~ accurate protein structure prediction by AlphaFold [1] and ~~other deep-learning-based methods~~RoseTTAFold [2]), it is now feasible to predict ~~static~~ native structures for many ~~many~~ proteins ~~of interest~~ given their sequences (with some caveats, see [3]), thus practically solving the protein folding problem [4]. However, it remains challenging to predict dynamic structural ensembles [5] and mutation-induced ~~structural changes~~effects [6] to meet the demand of mechanistic studies of protein functions and dysfunctions. While the public databases of protein sequences and variations increase rapidly owning to genomic/metagenomic sequencing efforts (~~e.g.~~ the MetaClust database contains about 1.6 billion protein sequence fragments [7]), the growth of experimental protein structures [8] and predicted structures remains to catch up ((~~e.g.~~ the AlphaFold database contains over 200 million predicted structures [9]). Such sequences-structures gap has motivated the development of new computational tools that make functional sense of protein sequences without directly using structural information (for example, by using deep learning to train large protein language models [10]). Recently, AlphaMissense attained state of the art prediction of missense variant pathogenicity by adapting AlphaFold fine-tuned on human and primate variant population frequency databases [11]. ~~.~~

A major interest in personalized ~~genetics and~~ medicine is in understanding novel genetic variations through genotype-phenotype association studies in relation to diseases. Particularly, a rapidly growing number of non-synonymous single nucleotide variants (nSNVs) have been

uncovered in protein coding regions that can adversely impact protein function and cause diseases [12]. Various computational methods were developed using evolutionary conservation and phylogeny in combination with biochemical and structural properties of amino acids to discriminate neutral vs. deleterious nSNVs [13-22]. Protein structural dynamics has also proven useful in discovering functionally important residues [23,24] which could constitute hot spots for disease-causing nSNVs [25,26]. However, the requirement of 3D structures has limited the number of nSNVs that can be analyzed by existing structure-based computational tools, although such constraint has been significantly alleviated by recent progress in protein structure prediction [27].

As alternatives to structure-based methods, sequences-based coevolution analysis has become increasingly powerful in predicting structural couplings between pairs of contacting residues [28-31] , thanks to the development of direct coupling methods that can overcome the confounding indirect coupling effects [29,32,33] . In principle, coevolving pairs of residues can be identified from a sufficiently large multiple sequence alignment, allowing the prediction of close spatial proximity in the native structures. Boosted by deep learning and other algorithmic developments, this coevolution analysis has led to accurate prediction of residue contacts which make *de novo* protein structure prediction possible [28] . Furthermore, coevolution analysis (enhanced by deep learning) has also been used to study various aspects of protein functional interactions such as allostery [34] . For example, RaptorX uses an ultra-deep neural network combining coevolution information with sequence conservation information to infer 3D contacts with higher accuracy than previous methods [35,36]. DeepMetaPSICOV [37] combines the input feature sets used by earlier methods (~~e.g.~~ MetaPSICOV [38] and DeepCov [39]) as input to a deep, fully convolutional residual neural network. SPOT-Contact predicts protein contact maps

by stacking residual convolutional networks with two-dimensional residual bidirectional recurrent LSTM networks, and using both one-dimensional sequence-based and two-dimensional evolutionary coupling based information [40]. These three state-of-the-art coevolution analysis methods are employed in this study to construct protein residue contact maps for network analysis (see below).

Another line of protein research is based on the treatment of a protein as a network where amino acid residues are nodes and their bonded/non-bonded interactions form edges [41]. Such models can be readily built upon 3D native structures so that a whole suite of network analysis tools (see https://networkx.org/) can be applied. For example, Amitai et al [42] used network analysis of protein structures (usinge.g. closeness centrality) to identify functional residues. Going beyond network analysis, deep-learning-based study of protein graph neural networks is an active area of research [43].

In a recent paper, Butler et al [44] proposed a sequence-based Gaussian network model (Seq-GNM) to calculate the dynamic profile of a protein without a 3D structure. They used coevolution analysis to build a network model which connects residues predicted to be in contact via evolutionary couplings. Their work built on previous studies that shown crystallographic B-factors are useful in predicting the impact of nSNVs on protein function [45,46] : rigid sites with low B-factors are more susceptible to destabilizing nSNVs than flexible sites with high B-factors. Indeed, existing computational tools to diagnose neutral and deleterious nSNVs (e.g.such as PolyPhen-2 [47]) use crystallographic B-factors along with other evolutionary and structural features. More specifically, Butler et al used Seq-GNM to compute B-factors for protein residues, and they found that deleterious nSNVs are overabundant at low B-factor sites, while neutral nSNVs are overabundant at high B-factor sites. Mechanistically, low B-factors may

indicate that a site is crucial for maintaining structural stability and/or modulating functional motions (~~e.g.,~~ as a hinge) and thus susceptible to mutations. In contrast, high B-factors are associated with flexible regions ~~(e.g., loops)~~ with minimal interactions, which are thus more robust to mutations. Based on these observations, they proposed that the sequences-based predicted B-factors can discriminate between deleterious and neutral nSNVs without structural information.

Inspired by the above study and recent progress in machine learning, network analysis, and protein language models, we further develop the sequences-based protein residue network analysis in the following directions: 1. We ~~exploit and integrate various methods of~~ build~~ing~~ protein residue networks using three ~~state-of-the-art~~different coevolution analysis tools (~~e.g.,~~ RaptorX, DeepMetaPSICOV, and SPOT-Contact) as ~~powered~~ enabled by deep learning. 2. We ~~use~~ exploit three machine learning algorithms (~~e.g.,~~ Random Forest, Gradient Boosting, and Extreme Gradient Boosting) to optimally combine ~~13~~ 20 distinct network node centrality scores ~~(calculated by the NetworkX package) with~~as ~~7 other network scores~~ calculated from the contact probability matrices to ~~jointly~~ predict ~~key residues as~~ hot spot residue~~s~~ for disease mutations. 3. ~~Using~~ Based on a dataset of 107 proteins ~~rich in~~with known ~~disease~~ deleterious/neutral mutations, we ~~rigorously~~ evaluate ~~the~~ our sequences-based ~~sequences-based~~ network scores both individually and in combination, and then ~~(via machine learning) in~~ compar~~e~~ison with alternative structures-based network scores and a physics force field based method ~~(using predicted structures by AlphaFold)~~. By optimally combing three coevolution analysis methods and the resulting 20 network scores by machine learning, we are able to discriminate deleterious and neutral mutation sites accurately (AUC of ROC ~ 0.84), which is on par with structure-based network scores (AUC ~ 0.83). Furthermore, by combining our method with a state-of-the-art

predictor of the functional effects of sequence variation based on large protein language models (ESM [48]), we have significantly improved the prediction of disease variant sites (AUC ~ 0.89).

In the following sections, we first describe the detailed methodology in the order of the proposed workflow, then we report the results of evaluation of our network-based scores both individually and collectively (via machine learning), finally we discuss specific case studies of four proteins to illustrate the usage of our method.~~This work supports the strategy of combining an ensemble of network scores based on different coevolution analysis methods via machine learning to predict candidate sites for disease mutations which will inform many downstream biomedical applications.~~

Materials and methods

Here is a summary of the workflow of our sequences-based method:

a. Collect datasets of protein sequences and variants (see Section 1)
b. Run co-evolution analysis of a given target protein sequence to build a residue contact map P (see Section 2)
c. Use NetworkX to calculate node centrality scores based on P (see Section 3)
d. Use sequence-based GNM to calculate additional node scores (see Section 4)
e. (optional) Use protein language model (ESM) to predict variant importance (see Section 5)
f. (optional) Use AlphaFold and FoldX to predict variant importance (see Section 6 and 7)
g. Use machine learning to optimally combine the above scores for classifying deleterious vs neutral variant sites (see Section 8)

Formatted: List Paragraph, Numbered + Level: 1 + Numbering Style: a, b, c, … + Start at: 1 + Alignment: Left + Aligned at: 0.5" + Indent at: 0.75"

Formatted: Font: (Default) Times New Roman, 12 pt

Formatted: Font: 12 pt, Not Bold

1. ~~Training and testing~~ Ddatasets of protein sequences and variants

A dataset of 107 protein sequences with ≤500 residues and ≥20 annotated deleterious/neutral variants were collected from the HumVar database [47] (sources: humvar-2011_12.deleterious.pph.input and humvar-2011_12.neutral.pph.input from ftp://genetics.bwh.harvard.edu/pph2/training/training-2.2.2.tar.gz). Their UniProt ids and sequences~~are as follows: O14896,O15305,O15537,O43826,O60260,O60880,O60931,P00439,P00441,P00480, P00492,P00740,P00742,P00813,P00966,P01008,P01009,P01185,P01308,P01909,P01920,P02042,P02533,P02649,P02766,P04070,P04180,P04181,P04440,P05155,P06132,P06280,P06858,P07902~~ can be accessed at https://simtk.org/projects/hotspots. This diverse dataset contains 97 proteins with their pairwise sequence identity < 30%.

~~,P08034,P08100,P08246,P08397,P08559,P08686,P08709,P10746,P10828,P11166,P16930,P17302,P17661,P19429,P19544,P21549,P22830,P23760,P23942,P25189,P25445,P26367,P26439, P29033,P29965,P30518,P30566,P30793,P31213,P31785,P32245,P35557,P35575,P39019,P41181,P42771,P45379,P45381,P49768,P51608,P51648,P51681,P51810,P53634,P56539,P58012,P60201,P63092,P68133,P68871,P69891,P69892,P69905,P82251,P98172,Q00604,Q01453,Q03393, Q03426,Q05066,Q13148,Q14654,Q15465,Q16586,Q6VVB1,Q92838,Q92947,Q93099,Q96NR8,Q99519,Q9GZX3,Q9H9S5,Q9UBM7.~~

The HumVar dataset consists of 13,032 human disease-causing mutations from UniProt and 8,946 human nonsynonymous single-nucleotide polymorphisms (nsSNPs) without annotated involvement in disease. This dataset was previously used to train and test PolyPhen-2 [47] for predicting damaging effects of missense mutations, and was used by Butler et al [44] in

Formatted: Font color: Auto

Formatted: Font color: Auto

Formatted: Font color: Auto

benchmarking their seq-GNM method for predicting deleterious/neutral nSNVs. ~~However, because they used different subsets of HumVar to train and test their methods, it is not possible to directly compare the performance between our method and theirs.~~

Since this dataset is highly imbalanced (there are 4040 deleterious mutation sites but only 120 neutral mutation sites) [49], we have added 3403 additional neutral sites with very low conservation scores (i.e. grade ≤2 as assessed by the ConSurf program [50]). Our objective is to train and test a binary classifier of residues in these proteins as deleterious or neutral. To this end, we split 107 proteins into training and testing sets (with 79 and 28 proteins, respectively), and perform evaluations based on the testing set. The main metric of evaluation is the ROC curves and associated area under the curve (AUC). AUC is a standard metric for evaluating binary classifiers based on the ROC curve of sensitivity and specificity. The ROC curves are also used in other computational papers for variant prediction (see [47]).

Formatted: Font: (Default) Times New Roman, 12 pt, Font color: Auto

Formatted: Font: (Default) Times New Roman, 12 pt, Font color: Auto

Formatted: Font: (Default) Times New Roman, 12 pt, Font color: Auto

~~S~~2. Sequences-based coevolution analysis and protein contact map construction

We perform coevolution analysis using three state-of-the-art methods: the RaptorX server (http://raptorx.uchicago.edu), the DeepMetaPSICOV server (http://bioinf.cs.ucl.ac.uk/psipred/), and the SPOT-Contact server (https://sparks-lab.org/server/spot-contact/). A sequence length limit (500) is imposed by the capacity of coevolution analysis servers, and may be circumvented if installing and running coevolution analysis locally.

These methods use multiple sequence alignments to compute the probability P_{ij} of residue pair (i, j) forming spatial contact. Based on the matrix of predicted P_{ij}, a protein residue contact map can be built with residues as nodes and pairwise contacts as edges weighted by P_{ij}. By default, we do not apply any threshold cutoff to P_{ij} for defining contacts (unless networks with unweighted edges are required by some node centrality algorithms in NetworkX, where we remove edges with $P_{ij}<0.1$, and set weight to 1 for the remaining edges).

3. Network analysis of protein contact map

By treating a protein contact map as a network of nodes and edges, we calculate various node centrality scores to predict key residues as hotspots for disease mutations.

A simple score to measure node centrality is a weighted node degree that accounts for the nearest neighbor interactions (denoted W_1):

$$W_{1,i} = \sum_{k \neq i} P_{ik} \qquad (1)$$

To include indirect couplings beyond the nearest neighbors, we calculate the node degree based on the n'th power of the contact probability matrix (denoted W_n):

$$W_{n,i} = \sum_{k \neq i} P_{ik} W_{n-1,k} = \sum_{k \neq i} P^n{}_{ik} \qquad (2)$$

As n goes to infinity, W_n converges to the eigenvector of P matrix with the highest eigenvalue $\lambda_{\max}$ (denoted W_∞):

$$PW_\infty = \lambda_{\max} W_\infty \quad (3)$$

Among various W_n, W_2 can be interpreted as the node degrees of a new network based on a neighborhood similarity matrix S as follows (denoted W_s):

$$S_{ij} = \sum_{k \neq i,j} P_{ik} P_{jk}, \; W_{s,i} = \sum_{k \neq i} S_{ik} \quad (4)$$

In this study we use five network scores (W_1, W_2, W_3, W_∞ and W_s) as predictive features for node importance. Additionally, we exploit 13 network centrality metrics as calculated by the NetworkX package (see Table 1). To allow meaningful comparison of scores between proteins, the scores of each protein are sorted and their ranking percentiles are linearly transformed to values between 0 and 1.

Table 1. Network centrality scores as implemented in the NetworkX package

(see https://networkx.org/documentation/stable/reference/algorithms/centrality.html)

Symbol	Centrality name	Definition
C1	degree_centrality	Corresponding to W_1
C2	eigenvector_centrality	Corresponding to W_∞
C3	closeness_centrality	Closeness centrality of a node u is the reciprocal of the average shortest path distance to u over all n-1 reachable nodes.
C4	betweenness_centrality	Betweenness centrality of a node u is the sum of the fraction of all-pairs shortest paths that pass through u.
C5	current_flow_closeness_centrality	Current-flow closeness centrality is a variant of closeness centrality based on effective resistance between nodes in a network.
C6	current_flow_betweenness_centrality	Current-flow betweenness centrality is based on an electrical current model for information spreading.
C7	communicability_betweenness_centrality	Communicability betweenness centrality is based on the number of walks connecting every pair of nodes.
C8	load_centrality	Load centrality of a node u is the fraction of all shortest paths that pass through u.
C9	subgraph_centrality	Subgraph centrality of a node u is the sum of weighted closed walks of all lengths starting and ending at u.
C10	harmonic_centrality	Harmonic centrality of a node u is the sum of the reciprocal of the shortest path distances from all other nodes to u.

C11	second_order_centrality	Second order centrality of a node u is the standard deviation of the return times to u of a perpetual random walk on G.
C12	laplacian_centrality	Laplacian Centrality of a node u is measured by the drop in the Laplacian Energy after deleting u from the graph.
C13	katz_centrality_numpy	Katz centrality computes the centrality for a node u based on the centrality of its neighbors. It is a generalization of the eigenvector centrality.

4. Sequences-based GNM

For comparison, we implemented Bulter et al's sequence-based GNM [44]. The original structure-based Gaussian network model (GNM) represents a protein structure as an elastically connected network of residues to obtain the equilibrium fluctuations of residues. In the absence of a structure, the sequence-based GNM (Seq-GNM) treats coevolving residue pairs as contacting pairs.

To construct the Kirchhoff matrix (denoted *K*), each non-bonded residue pair is assigned a value of -1 times its contact probability. The bonded residue pairs (i, i+1) are assigned -1 to enforce local chain connectivity. The diagonal elements of *K* are assigned so that the sum of each row and column is zero:

$$K_{ij} = \begin{cases} -P_{ij} & i \neq j \\ \sum_{k \neq i} P_{ik} & i = j \end{cases} \quad (5)$$

The vibrational thermal fluctuations of residues are evaluated by inverting the Kirchhoff matrix (or summing over the modes as weighted by $1/\lambda_m$). The per-residue mean-square

fluctuations (MSF), which are proportional to the crystallographic B factors, are given as follows:

$$MSF_i \propto K_{ii}^{-1} = \sum_{m>0} \frac{V_{mi}^{\ 2}}{\lambda_m} \quad (6)$$

where the eigen-decomposition of K gives eigenvectors V_m and eigenvalues λ_m that satisfy:

$$KV_m = \lambda_m V_m \quad (7)$$

Low-MSF residues correspond to rigid cores or hinges of dynamical importance [44].

As an alternative way to evaluate node importance using GNM, we perform a perturbation-based hotspot analysis as follows: For mode m, calculate how much its eigenvalue changes ($\delta\lambda_{m,i}$) in response to a perturbation at a chosen residue position i [23,24,51] (i.e., by uniformly weakening the contacts with residue i). Then compute $\delta\lambda_i = \sum_m \delta\lambda_{m,i}$ to assess the dynamic importance of this residue position [52]. High-$\delta\lambda_i$ residues correspond to sites highly sensitive to local perturbations that mimic mutations.

The above two GNM-based scores are combined with the other network scores for machine learning.

5. ESM based variant prediction

For comparison with our method, we use a deep-learning variant predictor based on a large protein language model (ESM). We downloaded and installed the ESM package and

pretrained models from https://github.com/facebookresearch/esm. Since our dataset consists of known variants (from HumVar) and added non-conserved sites (with specific mutations unknown), we simulate the mutational effects on each site by introducing Alanine substitution if the wildtype residue is not an Alanine and Glycine substitution otherwise. Then we process the mutated sequence with 5 pretrained ESM models (esm1v_t33_650M_UR90S_1, esm1v_t33_650M_UR90S_2, esm1v_t33_650M_UR90S_3, esm1v_t33_650M_UR90S_4, and esm1v_t33_650M_UR90S_5), which predict the difference in the probability of observing the widetype residue and the mutant residue at a given site [48]. We record the predictions of five ESM models as separate features to be optimally integrated via machine learning.

6. AlphaFold for structural prediction

We downloaded predicted structures for the 107 proteins from AlphaFold DB (https://alphafold.ebi.ac.uk/). A residue contact probability matrix is constructed based on the predicted structures as follows:

$$P_{ij} = \frac{1}{1+e^{d_{ij}-10}} \quad (8)$$

where d_{ij} is the distance between residues i and j, and 10 Å is used as a soft cutoff distance. We then use this contact probability matrix to perform the same network analysis as in the sequences-based method and for optimization with machine learning.

7. Foldx for structural refinement and Alanine scanning analysis

FoldX program [53] was downloaded from https://foldxsuite.crg.eu/. We use the RepairPDB command to refine the AlphaFold-predicted models (by fixing bad torsion angles and Van der Waals clashes). Then we use the AlaScan command to mutate each residue to Ala and calculate the resulting changes in Gibbs free energies which are then used as a feature to predict hotspots of disease mutations.

8. Machine learning algorithms

We use the following machine learning methods of the scikit-learn package (https://scikit-learn.org/stable/) to learn optimal combinations of all features to predict if a given site is deleterious or neutral mutation site:

Random Forest Classifier (RF) (sklearn.ensemble.RandomForestClassifier): A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. We tune the following hyper-parameters: max_depth, n_estimators, max_features.

Gradient Boosting Classifier (GB) (sklearn.ensemble.GradientBoostingClassifier): This algorithm builds an additive model in a forward stage-wise fashion. In each stage a regression tree is fit on the negative gradient of the loss function, e.g. binary log loss. We tune the following hyper-parameters: n_estimators, max_depth, max_features.

- Extreme Gradient Boosting Classifier (XGB) (xgboost.XGBClassifier): This algorithm is an optimized distributed version of gradient boosting designed to be highly efficient, flexible and

portable. We tune the following hyper-parameters: n_estimators, max_depth, reg_alpha, reg_lambda.

These three methods were chosen because they have performed successfully in machine learning contests in Kaggle (see https://www.packtpub.com/product/the-kaggle-book/9781801817479). They are also relatively cheap to train and optimize compared with the deep learning methods.

We use Optuna (https://optuna.org/) for hyper-parameter tuning of the above algorithms. We have run Optuna multiple times to ensure the resulting best metric is reproducible.

Formatted: Font color: Auto

Results and discussion

This study explores how to systematically utilize the coevolution information from multiple sequence alignments to model and analyze a protein as a residue contact network beyond the scope of GNM. To this end, we first use coevolution analysis to construct a protein residue contact map with edges weighted by the predicted contact probability; then we exploit an array of 20 network-based scores to assess the node importance as predictors for disease mutation sites; finally we evaluate the predictive power of these scores individually and collectively (using machine learning) based on a subset of 107 protein sequences and their variants from the HumVar database. For comparison, we also evaluate alternative methods based on predicted protein structures, a physics-based force field, and protein language models.

Formatted: Indent: First line: 0"

1.

Evaluation of individual network scores

Based on the protein residue contact maps built from three coevolution analysis tools (DeepMetaPSICOV, RaptorX, and SPOT-Contact), we applied network analysis to calculate 20 network scores (see Table 2), measuring node centrality using various different algorithms (see Methods). These scores include simple weighted node degrees for n-hop nearest neighbors (see Methods) and more sophisticated centrality metrics (see Table 1), along with 2 seq-GNM based scores (MSF and δλ, see Methods). We evaluate the performance of each score using the AUC of ROC for the testing set, which provides a balanced evaluation of sensitivity and specificity ~~as~~

~~functions of the cutoff score~~ (see Table 2). More specifically, we sort all testing-set variants by a particular score and predict a variant deleterious/neutral if its score is above/below a cutoff value. This results in an ROC curve from which we have calculated AUC (see Table 2).

Formatted: Font color: Auto

Formatted: Font color: Auto

Formatted: Font color: Auto

Formatted: Font color: Auto

Overall, DeepMetaPSICOV (max AUC=0.80) and SPOT-Contact (max AUC=0.81) perform slightly better than RaptorX (max AUC=0.78). Interestingly, simple weighted node degrees (W_1, W_2, and W_3) perform better than those more complex centrality scores (see Table 2). When computing node degrees, going beyond the nearest neighbors seems to improve the prediction slightly (see Table 2). Two GNM-based scores perform similarly but slightly worse than the weighted node degrees (see Table 2). Among those NetworkX-based scores (see Table 1), C5, C11 and C12 outperform the others, while those betweenness-based scores (C4, C6, and C8) underperform (see Table 2). Therefore, the functional importance of a node/residue pertains more to its role as a highly-connected hub than as an information bottleneck of the shortest paths.

Table 2. Evaluation of 20 network scores based on protein residue contact maps constructed from 3 coevolution analysis tools (DeepMetaPSICOV, RaptorX, and SPOT-Contact) and AlphaFold-predicted structures

Score	AUC* of DeepMetaPSICOV	AUC* of RaptorX	AUC* of SPOT-Contact	AUC* of AlphaFold
C1	0.74	0.76	0.73	0.82
C2	0.73	0.74	0.76	0.77
C3	0.76	0.73	0.69	0.73
C4	0.64	0.54	0.60	0.58
C5	0.78	0.76	0.79	0.80
C6	0.63	0.58	0.67	0.60
C7	0.75	0.61	0.72	0.74
C8	0.64	0.54	0.60	0.58
C9	0.77	0.76	0.74	0.78
C10	0.75	0.73	0.68	0.75
C11	0.79	0.76	0.77	0.80
C12	0.77	0.77	0.79	0.83
C13	0.73	0.73	0.76	0.76

$\delta\lambda$	0.79	0.76	0.78	0.83
MSF	0.79	0.76	0.78	0.80
W_1	0.79	0.77	0.78	**0.83**
W_2	0.80	0.78	0.80	0.83
W_3	**0.80**	**0.78**	**0.81**	0.82
W_∞	0.80	0.74	0.79	0.77
W_s	0.80	0.78	0.80	0.83
FoldX				0.68

* The AUC is calculated based on the ROC for all variants of the 28 testing set proteins. Alternatively, we also calculated AUCs based on the ROCs of individual proteins and their summary statistics (see Table S1).

For comparison with alternative methods, we evaluated the performance of variant prediction by five pre-trained protein language models (ESM, see Methods), and the resulting AUC varies between 0.79 and 0.81, which are comparable to the network scores (see Table 2). For further comparison with structures-based methods, we also performed network analysis based on protein structures as predicted by AlphaFold (see Methods). Overall, the structures-based scores (max AUC=0.83) perform slightly better than the sequences-based scores. This may be partly due to the structure-based contact maps (see Eq. 8) being more sharply defined than the fuzzier contact-probability-based contact maps. Notably, when structural information is used, our network analysis performs significantly better than a physics-based force field (FoldX) with AUC=0.68. Taken together, these findings support the usefulness of individual sequences-based network centrality scores in predicting important residues on par with alternative more sophisticated methods.

To further understand the different accuracies of the above scores, we explore the relationships between them by evaluating the pairwise Pearson correlations (PC) (see Table 3). W_1, W_2, W_3, W_∞, W_s , MSF and $\delta\lambda$ are highly correlated (with PC≥0.93 for DeepMetaPSICOV,

PC≥0.84 for SPOT-Contact, PC≥0.86 for AlphaFold), although their correlations are somewhat weaker for RaptorX. Among the NetworkX-based scores (see Table 1), C5, C11 and C12 are also highly correlated with the above scores. Such strong correlations support the attribution of higher AUC of these scores (see Table 2) to their capturing the same essential features (i.e. high node degrees) of those important nodes. In contrast, the betweenness-based scores (C4, C6, and C8) do not correlate well with the above scores, which is consistent with their lower AUC (see Table 2).

Table 3. Pearson correlations between network scores (row 1, 2, 3 and 4 correspond to results of DeepMetaPSICOV, SPOT-Contact, RaptorX and AlphaFold, respectively)

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	W_1	W_2	W_3	W_∞	W_s	MSF	$\delta\lambda$
C1	1.00	0.57	0.74	0.63	0.72	0.40	0.82	0.63	0.84	0.79	0.88	0.84	0.54	0.87	0.85	0.83	0.80	0.85	0.87	0.87
C2	0.57	1.00	0.43	0.21	0.75	0.18	0.47	0.21	0.66	0.42	0.67	0.76	0.98	0.65	0.72	0.75	0.80	0.71	0.65	0.65
C3	0.74	0.43	1.00	0.64	0.63	0.36	0.72	0.64	0.73	0.96	0.81	0.57	0.43	0.63	0.63	0.63	0.61	0.64	0.64	0.63
C4	0.63	0.21	0.64	1.00	0.34	0.42	0.64	1.00	0.40	0.65	0.53	0.37	0.21	0.46	0.42	0.40	0.38	0.43	0.47	0.46
C5	0.72	0.75	0.63	0.34	1.00	0.46	0.67	0.34	0.74	0.58	**0.91**	**0.86**	0.76	**0.80**	**0.83**	**0.84**	**0.84**	**0.83**	**0.80**	**0.80**
C6	0.40	0.18	0.36	0.42	0.46	1.00	0.63	0.42	0.21	0.29	0.45	0.33	0.20	0.45	0.41	0.38	0.34	0.42	0.46	0.45
C7	0.82	0.47	0.72	0.64	0.67	0.63	1.00	0.64	0.73	0.71	0.80	0.68	0.46	0.76	0.75	0.74	0.71	0.76	0.77	0.76
C8	0.63	0.21	0.64	1.00	0.34	0.42	0.64	1.00	0.40	0.65	0.53	0.37	0.21	0.46	0.42	0.40	0.37	0.43	0.47	0.46
C9	0.84	0.66	0.73	0.40	0.74	0.21	0.73	0.40	1.00	0.79	0.85	0.79	0.63	0.77	0.81	0.82	0.82	0.81	0.78	0.77
C10	0.79	0.42	0.96	0.65	0.58	0.29	0.71	0.65	0.79	1.00	0.79	0.59	0.41	0.65	0.65	0.64	0.63	0.66	0.66	0.65
C11	0.88	0.67	0.81	0.53	0.91	0.45	0.80	0.53	0.85	0.79	1.00	**0.85**	0.68	**0.84**	**0.85**	**0.85**	**0.84**	**0.86**	**0.85**	**0.84**
C12	0.84	0.76	0.57	0.37	0.86	0.33	0.68	0.37	0.79	0.59	**0.85**	1.00	0.74	**0.91**	**0.94**	**0.94**	**0.92**	**0.92**	**0.90**	**0.91**
C13	0.54	0.98	0.43	0.21	0.76	0.20	0.46	0.21	0.63	0.41	0.68	0.74	1.00	0.63	0.69	0.73	0.78	0.69	0.63	0.63
W_1	0.87	0.65	0.63	0.46	0.80	0.45	0.76	0.46	0.77	0.65	0.84	0.91	0.63	**1.00**	**0.98**	**0.97**	**0.93**	**0.98**	**1.00**	**1.00**
W_2	0.85	0.72	0.63	0.42	0.83	0.41	0.75	0.42	0.81	0.65	0.85	0.94	0.69	**0.98**	**1.00**	**1.00**	**0.97**	**1.00**	**0.99**	**0.98**
W_3	0.83	0.75	0.63	0.40	0.84	0.38	0.74	0.40	0.82	0.64	0.85	0.94	0.73	**0.97**	**1.00**	**1.00**	**0.99**	**0.99**	**0.97**	**0.97**
W_∞	0.80	0.80	0.61	0.38	0.84	0.34	0.71	0.37	0.82	0.63	0.84	0.92	0.78	**0.93**	**0.97**	**0.99**	**1.00**	**0.97**	**0.93**	**0.93**
W_s	0.85	0.71	0.64	0.43	0.83	0.42	0.76	0.43	0.81	0.66	0.86	0.92	0.69	**0.98**	**1.00**	**0.99**	**0.97**	**1.00**	**0.99**	**0.98**
MSF	0.87	0.65	0.64	0.47	0.80	0.46	0.77	0.47	0.78	0.66	0.85	0.90	0.63	**1.00**	**0.99**	**0.97**	**0.93**	**0.99**	**1.00**	**1.00**
C1	1.00	0.58	0.68	0.55	0.73	0.53	0.85	0.55	0.86	0.73	0.90	0.74	0.57	0.80	0.78	0.77	0.67	0.80	0.82	0.80
C2	0.58	1.00	0.44	0.20	0.88	0.40	0.51	0.20	0.61	0.38	0.74	0.82	0.99	0.70	0.76	0.79	0.88	0.74	0.68	0.70
C3	0.68	0.44	1.00	0.61	0.55	0.38	0.69	0.61	0.73	0.96	0.78	0.42	0.45	0.50	0.52	0.53	0.52	0.55	0.56	0.50
C4	0.55	0.20	0.61	1.00	0.29	0.46	0.63	1.00	0.37	0.62	0.49	0.26	0.20	0.37	0.35	0.34	0.30	0.37	0.40	0.37
C5	0.73	0.88	0.55	0.29	1.00	0.57	0.67	0.29	0.71	0.49	**0.88**	**0.91**	**0.87**	**0.84**	**0.88**	**0.89**	**0.87**	**0.87**	**0.82**	**0.84**
C6	0.53	0.40	0.38	0.46	0.57	1.00	0.73	0.46	0.33	0.31	0.54	0.55	0.40	0.64	0.61	0.59	0.49	0.62	0.66	0.64
C7	0.85	0.51	0.69	0.63	0.67	0.73	1.00	0.63	0.75	0.68	0.82	0.66	0.51	0.76	0.75	0.73	0.64	0.77	0.80	0.76
C8	0.55	0.20	0.61	1.00	0.29	0.46	0.63	1.00	0.37	0.62	0.49	0.26	0.20	0.37	0.35	0.34	0.30	0.37	0.40	0.37
C9	0.86	0.61	0.73	0.37	0.71	0.33	0.75	0.37	1.00	0.77	0.87	0.66	0.59	0.68	0.71	0.72	0.68	0.73	0.71	0.68
C10	0.73	0.38	0.96	0.62	0.49	0.31	0.68	0.62	0.77	1.00	0.76	0.38	0.38	0.47	0.49	0.49	0.47	0.52	0.54	0.47
C11	0.90	0.74	0.78	0.49	0.88	0.54	0.82	0.49	0.87	0.76	1.00	**0.79**	0.73	**0.80**	**0.82**	**0.83**	**0.78**	**0.83**	**0.82**	**0.80**
C12	0.74	0.82	0.42	0.26	0.91	0.55	0.66	0.26	0.66	0.38	**0.79**	1.00	0.80	**0.91**	**0.93**	**0.92**	**0.83**	**0.90**	**0.86**	**0.91**

C13	0.57 0.99 0.45 0.20 0.87 0.40 0.51 0.20 0.59 0.38 0.73 0.80 1.00 0.69 0.75 0.78 0.88 0.73 0.67 0.69
W_1	0.80 0.70 0.50 0.37 0.84 0.64 0.76 0.37 0.68 0.47 0.80 0.91 0.69 **1.00 0.98 0.97 0.84 0.98 0.98 1.00**
W_2	0.78 0.76 0.52 0.35 0.88 0.61 0.75 0.35 0.71 0.49 0.82 0.93 0.75 **0.98 1.00 1.00 0.90 1.00 0.97 0.98**
W_3	0.77 0.79 0.53 0.34 0.89 0.59 0.73 0.34 0.72 0.49 0.83 0.92 0.78 **0.97 1.00 1.00 0.92 0.99 0.96 0.97**
W_∞	0.67 0.88 0.52 0.30 0.87 0.49 0.64 0.30 0.68 0.47 0.78 0.83 0.88 **0.84 0.90 0.92 1.00 0.90 0.86 0.84**
W_s	0.80 0.74 0.55 0.37 0.87 0.62 0.77 0.37 0.73 0.52 0.83 0.90 0.73 **0.98 1.00 0.99 0.90 1.00 0.98 0.98**
MSF	0.82 0.68 0.56 0.40 0.82 0.66 0.80 0.40 0.71 0.54 0.82 0.86 0.67 **0.98 0.97 0.96 0.86 0.98 1.00 0.98**
C1	1.00 0.63 0.58 0.22 0.64 0.22 0.29 0.22 0.75 0.65 0.71 0.78 0.55 0.88 0.87 0.85 0.58 0.87 0.68 0.73
C2	0.63 1.00 0.64 0.15 0.82 0.23 0.24 0.15 0.66 0.66 0.78 0.70 0.93 0.60 0.65 0.67 0.82 0.64 0.70 0.55
C3	0.58 0.64 1.00 0.44 0.79 0.35 0.39 0.44 0.64 0.93 0.88 0.51 0.68 0.48 0.50 0.50 0.64 0.50 0.69 0.47
C4	0.22 0.15 0.44 1.00 0.32 0.70 0.67 1.00 0.11 0.38 0.39 0.06 0.22 0.18 0.14 0.14 0.30 0.15 0.37 0.17
C5	0.64 0.82 0.79 0.32 1.00 0.45 0.43 0.32 0.65 0.72 **0.96** 0.71 **0.86** 0.63 0.65 0.66 0.79 0.65 **0.81** 0.62
C6	0.22 0.23 0.35 0.70 0.45 1.00 0.70 0.70 0.12 0.25 0.45 0.17 0.29 0.32 0.28 0.26 0.40 0.28 0.51 0.31
C7	0.29 0.24 0.39 0.67 0.43 0.70 1.00 0.67 0.34 0.34 0.48 0.28 0.29 0.30 0.28 0.27 0.40 0.29 0.49 0.39
C8	0.22 0.15 0.44 1.00 0.32 0.70 0.67 1.00 0.11 0.38 0.39 0.06 0.22 0.18 0.14 0.14 0.30 0.15 0.37 0.17
C9	0.75 0.66 0.64 0.11 0.65 0.12 0.34 0.11 1.00 0.71 0.72 0.71 0.61 0.66 0.69 0.69 0.62 0.69 0.66 0.64
C10	0.65 0.66 0.93 0.38 0.72 0.25 0.34 0.38 0.71 1.00 0.83 0.55 0.67 0.52 0.53 0.53 0.66 0.54 0.72 0.51
C11	0.71 0.78 0.88 0.39 0.96 0.45 0.48 0.39 0.72 0.83 1.00 0.68 **0.82** 0.64 0.65 0.65 0.77 0.65 **0.83** 0.63
C12	0.78 0.70 0.51 0.06 0.71 0.17 0.28 0.06 0.71 0.55 0.68 1.00 0.64 0.76 0.77 0.77 0.63 0.77 0.67 0.70
C13	0.55 0.93 0.68 0.22 0.86 0.29 0.29 0.22 0.61 0.67 0.82 0.64 1.00 0.53 0.58 0.61 0.85 0.58 0.72 0.52
W_1	0.88 0.60 0.48 0.18 0.63 0.32 0.30 0.18 0.66 0.52 0.64 0.76 0.53 **1.00 0.98 0.97** 0.69 **0.98 0.81 0.85**
W_2	0.87 0.65 0.50 0.14 0.65 0.28 0.28 0.14 0.69 0.53 0.65 0.77 0.58 **0.98 1.00 0.99** 0.73 **1.00 0.80 0.83**
W_3	0.85 0.67 0.50 0.14 0.66 0.26 0.27 0.14 0.69 0.53 0.65 0.77 0.61 **0.97 0.99 1.00** 0.75 **0.99 0.80 0.82**
W_∞	0.58 0.82 0.64 0.30 0.79 0.40 0.40 0.30 0.62 0.66 0.77 0.63 0.85 0.69 0.73 0.75 1.00 0.73 **0.88** 0.67
W_s	0.87 0.64 0.50 0.15 0.65 0.28 0.29 0.15 0.69 0.54 0.65 0.77 0.58 **0.98 1.00 0.99** 0.73 1.00 **0.81 0.83**
MSF	0.68 0.70 0.69 0.37 0.81 0.51 0.49 0.37 0.66 0.72 0.83 0.67 0.72 **0.81 0.80 0.80 0.88 0.81** 1.00 0.79
C1	1.00 0.87 0.77 0.38 0.95 0.42 0.82 0.38 0.90 0.81 0.96 0.98 0.85 0.97 0.97 0.96 0.86 0.97 0.95 0.97
C2	0.87 1.00 0.78 0.26 0.91 0.32 0.81 0.26 0.98 0.79 0.90 0.89 0.99 0.87 0.91 0.93 0.99 0.91 0.91 0.87
C3	0.77 0.78 1.00 0.52 0.79 0.37 0.74 0.52 0.82 0.97 0.83 0.72 0.78 0.71 0.74 0.75 0.78 0.74 0.79 0.71
C4	0.38 0.26 0.52 1.00 0.33 0.61 0.55 1.00 0.30 0.51 0.39 0.30 0.27 0.30 0.30 0.29 0.27 0.30 0.33 0.30
C5	0.95 0.91 0.79 0.33 1.00 0.46 0.85 0.33 0.92 0.79 **0.99 0.96** 0.91 **0.95 0.96 0.96 0.91 0.96 1.00 0.95**
C6	0.42 0.32 0.37 0.61 0.46 1.00 0.67 0.61 0.33 0.31 0.46 0.40 0.33 0.41 0.39 0.37 0.33 0.39 0.46 0.41
C7	0.82 0.81 0.74 0.55 0.85 0.67 1.00 0.55 0.83 0.72 0.85 0.81 0.81 0.80 0.82 0.83 0.81 0.82 0.84 0.80
C8	0.38 0.26 0.52 1.00 0.33 0.61 0.55 1.00 0.30 0.51 0.39 0.30 0.27 0.30 0.30 0.29 0.27 0.30 0.33 0.30
C9	0.90 0.98 0.82 0.30 0.92 0.33 0.83 0.30 1.00 0.84 0.93 0.91 0.96 0.89 0.93 0.95 0.97 0.93 0.92 0.89
C10	0.81 0.79 0.97 0.51 0.79 0.31 0.72 0.51 0.84 1.00 0.84 0.75 0.79 0.73 0.77 0.78 0.79 0.77 0.80 0.73
C11	0.96 0.90 0.83 0.39 0.99 0.46 0.85 0.39 0.93 0.84 1.00 **0.95** 0.90 **0.94 0.95 0.95 0.90 0.95 0.99 0.94**
C12	0.98 0.89 0.72 0.30 0.96 0.40 0.81 0.30 0.91 0.75 **0.95** 1.00 0.88 **1.00 1.00 0.99 0.89 0.99 0.96 1.00**
C13	0.85 0.99 0.78 0.27 0.91 0.33 0.81 0.27 0.96 0.79 0.90 0.88 1.00 0.85 0.90 0.92 0.99 0.90 0.91 0.85
W_1	0.97 0.87 0.71 0.30 0.95 0.41 0.80 0.30 0.89 0.73 0.94 1.00 0.85 **1.00 0.99 0.97 0.86 0.98 0.95 1.00**
W_2	0.97 0.91 0.74 0.30 0.96 0.39 0.82 0.30 0.93 0.77 0.95 1.00 0.90 **0.99 1.00 1.00 0.91 1.00 0.96 0.99**
W_3	0.96 0.93 0.75 0.29 0.96 0.37 0.83 0.29 0.95 0.78 0.95 0.99 0.92 **0.97 1.00 1.00 0.93 1.00 0.96 0.97**
W_∞	0.86 0.99 0.78 0.27 0.91 0.33 0.81 0.27 0.97 0.79 0.90 0.89 0.99 **0.86 0.91 0.93 1.00 0.91 0.91 0.86**
W_s	0.97 0.91 0.74 0.30 0.96 0.39 0.82 0.30 0.93 0.77 0.95 0.99 0.90 **0.98 1.00 1.00 0.91 1.00 0.96 0.98**
MSF	0.95 0.91 0.79 0.33 1.00 0.46 0.84 0.33 0.92 0.80 0.99 0.96 0.91 **0.95 0.96 0.96 0.91 0.96 1.00 0.95**

In summary, by evaluating 20 network scores individually, we have found a wide range of performance with AUC varying from 0.54 to 0.81 (see Table 2). The top-performing scores

seem to correlate strongly with each other, so they must have captured a common aspect of node centrality that is relevant to functional importance (e.g. high local connectivity instead of high betweenness). Interestingly, the two GNM-based scores, despite measuring distinct dynamic properties (MSF measures thermal fluctuations while $\delta\lambda$ measures sensitivity to local perturbations), are also strongly correlated with each other and those degree-based network scores. Therefore, to speed up the variant prediction workflow we only need to compute those simpler weighted node degrees as features without significantly losing accuracy.

2. Combining all network scores to predict variant hotspots by machine learning

To optimize the predictive power of the above network-based scores based on three coevolution analysis methods (or AlphaFold), we have employed machine learning algorithms (see Methods) to take them as input features, train a binary classifier which predicts if a residue position is linked to neutral or deleterious variants (using first 79 proteins as training set), and then test its prediction using the remaining 28 proteins as testing set. We use the AUC of ROC as the metric for assessing the prediction quality of the trained classifier.

To evaluate the protein residue contact maps constructed by each method, we combine all network scores based on the contact maps predicted by the same method (see Table 2) for machine learning. The resulting AUC of each coevolution analysis method (DeepMetaPSICOV, RaptorX, and SPOT-Contact) is 0.81, 0.80, and 0.82, respectively (see Table 4), which are slightly better than the best AUC of individual scores (0.78~0.81, see Table 2). The lack of substantial improvement may be due to high correlations among the scores (see Table 3) which

could reduce the effectiveness of ensemble learning. For comparison, we also trained and tested classifiers using the AlphaFold-predicted contact maps, and alternative classifiers based on protein language models (see Methods). Both alternative methods give comparable yet slightly better AUC (0.83). Similar to our finding, Butler et al reported AUC of 0.81 after combining the B-factors of Seq-GNM with evolutionary features [44].

Table 4. **Evaluation of classifiers trained by 3 machine learning algorithms (RF, GB and XGB, see Methods) based on the protein residue contact maps constructed from 3 coevolution analysis tools (DeepMetaPSICOV, RaptorX, and SPOT-Contact), AlphaFold-predicted structures, and protein language models (ESM).**

Sources of input features	AUC of RF	AUC of GB	AUC of XGB
DeepMetaPSICOV	0.81	0.81	0.81
RaptorX	0.80	0.80	0.80
SPOT-Contact	0.82	0.82	0.82
AlphaFold	0.83	0.83	0.83
ESM	0.83	0.83	0.83
All 3 coevolution methods	0.84	0.84	0.84
All 3 coevolution methods (w/o C1-C13)	0.82	0.82	0.83
All 3 coevolution methods and ESM	0.89	0.89	0.89
AlphaFold and ESM	0.88	0.88	0.88

To further boost the prediction performance, we have sought to combine the network scores of all three coevolution analysis methods for machine learning, resulting in better AUC (0.84) which slightly outperform both AlphaFold and ESM (0.83). To assess the added value of including 13 NetworkX-based centrality scores (see Table 1), we have performed an ablation study that excludes them in machine learning, and found slightly lower AUC (0.82~0.83). So it

is possible to speed up the calculation without significantly reducing accuracy. Taken together, our findings support the power of combining an array of different network scores from different coevolution analysis tools to optimize the prediction in the framework of ensemble learning.

To further explore how well our method complements alternative methods, we have combined all the network scores with the ESM scores in machine learning. Encouragingly, we have obtained markedly improved AUC (0.89), which is comparable to machine learning that combines the AlphaFold-based network scores with the ESM scores (AUC=0.88).

For comparison with other studies, Butler et al showed that Seq-GNM combined with evolutionary parameters attained a sensitivity of 0.84 and a specificity of 0.66 [44]. PolyPhen-2 achieved a sensitivity of 0.73 and a specificity of 0.8 on the HumVar datasets [47]. While using different training and testing datasets, we have attained competitive results with a sensitivity of 0.82 and a specificity of 0.80 (using all the network scores from three coevolution analysis tools and the ESM scores). For more direct comparison, we also evaluated PolyPhen-2 based on the same 28 testing-set proteins and their variants, and obtained an AUC of 0.85, which is close to our method (see Table 4). However, this metric is likely positively biased since PolyPhen-2 has been trained on the HumVar dataset.

In summary, via extensive machine learning, we have demonstrated the power of using an ensemble of sequences-based network scores calculated by different co-evolution analysis tools to accurately predict deleterious mutation sites. Although some network scores are highly correlated (see Table 3) and they vary widely in accuracy (see Table 2), these scores seem to be sufficiently diverse to allow effective ensemble learning when combined.

Formatted: Font: 12 pt, Font color: Auto

Formatted: Font: 12 pt, Font color: Auto

Formatted: Font: 12 pt, Font color: Auto

Formatted: Font color: Auto

3. Case studies:

To illustrate the biomedical significance of our predictions of variant sites with network scores, we discuss in details the following four proteins from our dataset.

Glucose-6-phosphate exchanger (Uniprot id: O43826): As an inorganic phosphate and glucose-6-phosphate antiporter, it transports cytoplasmic glucose-6-phosphate into the lumen of the endoplasmic reticulum and translocates inorganic phosphate in the opposite direction. Being involved in glucose production through glycogenolysis and gluconeogenesis, it plays a central role in homeostatic regulation of blood glucose levels. It is linked to diseases like congenital disorder of glycosylation and glycogen storage disease (see https://www.uniprot.org/uniprotkb/O43826/entry#function).

The AlphaFold-predicted structure forms a dimer of transmembrane helical domains with most deleterious mutation sites concentrated inside the central core while those non-conserved residues (i.e. neutral mutation sites) are mostly located on the periphery (see Fig 1c). The contact maps predicted by three coevolution analysis tools all agree well with the contact map based on the AlphaFold structure (see Fig 1a) (except that RaptorX omitted many local contacts in residues 1-200). As a result, the network centrality scores (~~e.g.~~ W_3) also agree well between these methods (see Fig 1b), although the coevolution-based network scores are generally noisier (with more spikes) than the structure-based scores (see Fig 1b). Different network scores calculated from the same contact map are also highly similar (see Fig 1d) despite being based on different algorithms. For example, scores of $\delta\lambda$ and MSF agree very well (see Fig 1d). Encouragingly, those residues identified with high network scores are primarily within the central core (inside each domain or in the inter-domain hinge region), thus overlapping with

most deleterious mutations (see Fig 1c). Among those top-10% hotspot residues (see Fig 1c), mutations Y24H, N27K, R28H, G88D, G149E, P153L, and G339C were implicated in causing glycogen storage disease [54] . Two of these mutations (R28H and G149E) were found to exhibit undetectable microsomal glucose-6-phosphate transport activity in transient expression studies[55], thus confirming their functional importance.

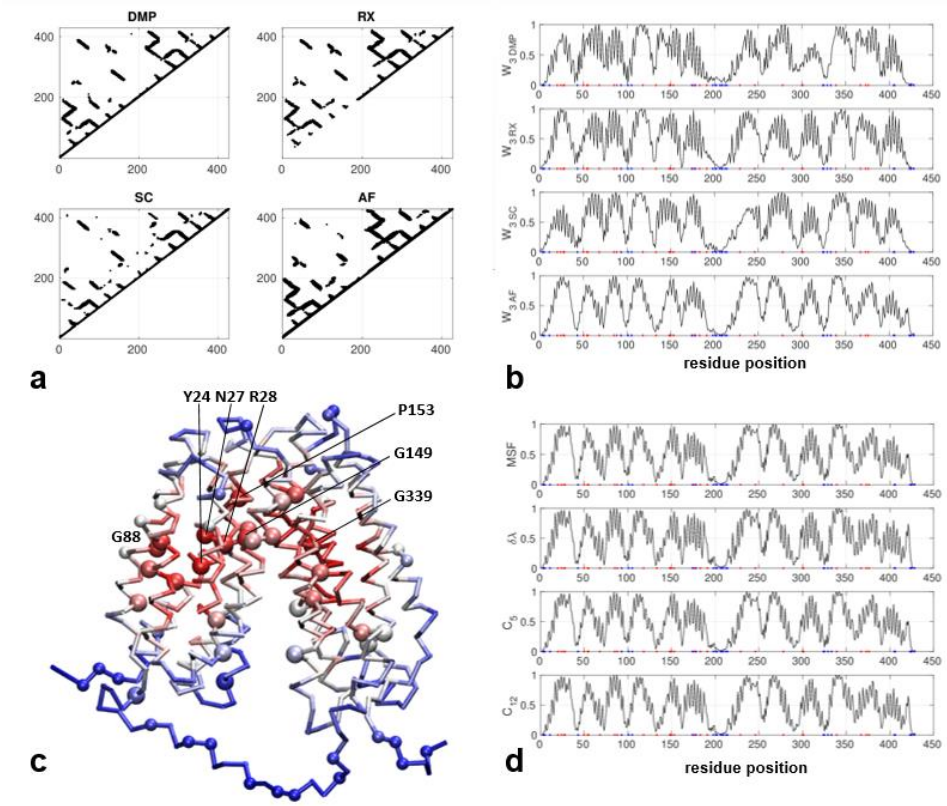


Figure 1. **Results for Glucose-6-phosphate exchanger (Uniprot id: O43826)**: (a) Four contact maps constructed from coevolution analysis by DeepMetaPSICOV (DMP), RaptorX (RX), SPOT-Contact (SC), and the predicted structure by AlphaFold (AF) (only those contacts with

probability >0.1 are shown). (b) W_3 scores for all residue positions based on the contact maps in (a), where red and blue dots mark residues with deleterious and neutral mutations, respectively. (c) Predicted structure by AlphaFold as colored by W_3 scores (red/blue for high/low values), where residues with deleterious and neutral mutations are shown as large and small balls, respectively (G20, Y24, N27, R28, G50, S54, S55, G68, L85, G88, W118, Q133, A148, G149, G150, P153, C176, C183, P191, L229, W246, I278, R300, H301, G339, A367, A373, G376, see https://www.uniprot.org/uniprotkb/O43826/variant-viewer). (d) Four other network scores (MSF, δλ, C5 and C12) for all residue positions based on the contact maps in (a).

Presenilin-1 (Uniprot id: P49768): As the catalytic subunit of the gamma-secretase complex, it catalyzes the intramembrane cleavage of integral membrane proteins such as Notch receptors. It is involved in various diseases including a familial early-onset form of Alzheimer disease and a form of frontotemporal dementia (see https://www.uniprot.org/uniprotkb/P49768/entry#function).

The AlphaFold-predicted structure consists of two closely packed helical domains with most deleterious mutations clustered inside the core domain while the non-conserved residues are mostly located on the N-terminal loop (residues 1-70) and the inter-domain linker (residues 300-370) (see Fig 2c). The active site [56] (D257 and D385) is also located in the core domain (colored green in Fig 2c). The contact maps predicted by three coevolution analysis methods all resemble the contact map based on the predicted structure (see Fig 2a) (except that RaptorX omitted local contacts in residues 1-100). As a result, the network scores agree well between them in the helical domains (see Fig 2c), but with more differences in the flexible regions (residues 1-70 and 300-370). Reassuringly, those residues identified by high network scores are

primarily clustered within the central core overlapping with most deleterious mutations, while the flexible N-terminal and linker feature low scores consistent with low sequence conservation (see Fig 2c). Among those top 10% hotspot residues (see Fig 2c), mutations at C92, V96, A231, M233, L235, A246, L250, S390, L392, and C410 were found to cause loss of function and altered amyloid-beta production [57] : C92S led to loss of protease function and increased Abeta42 levels. V96F caused loss of protease activity. A231T/V and M233T led to decreased protease activity, altered amyloid-beta production and increased amyloid-beta 42/amyloid-beta 40 ratio. L235P/R and S390I abolished protease activity. A246E and L250S abolished protease activity and increased amyloid-beta 42/amyloid-beta 40 ratio. L392V resulted in reduced proteolysis, altered amyloid-beta production and increased amyloid-beta 42/amyloid-beta 40 ratio. C410I reduced proteolysis. Since most of these residues are not near the active site, their effects on protease activity are likely allosteric.

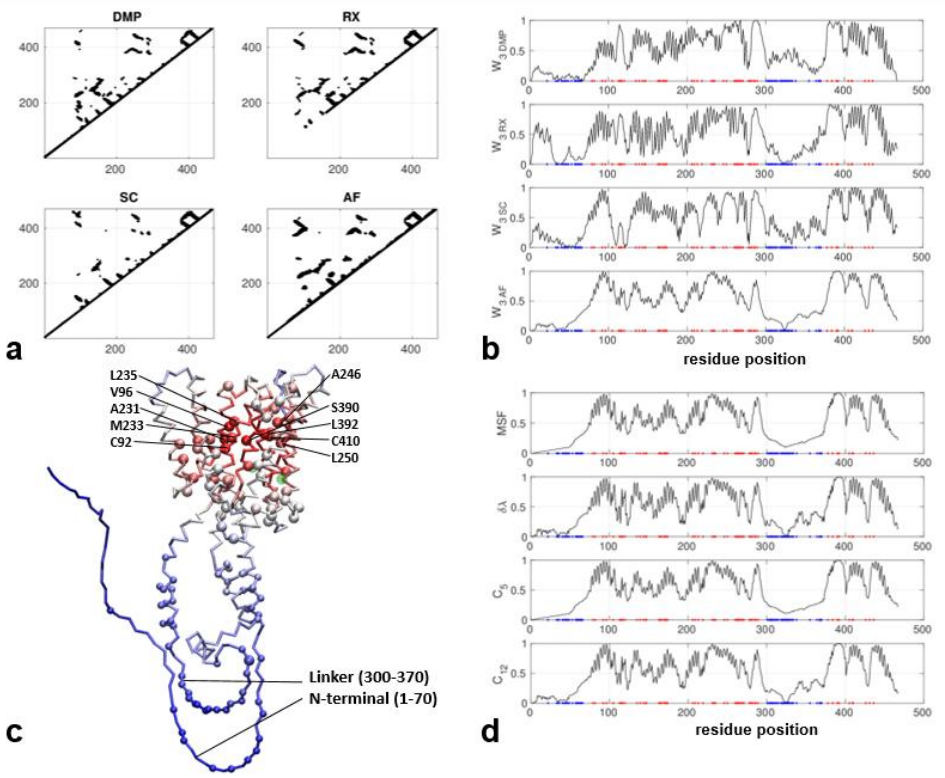


Figure 2. **Results for Presenilin-1 (Uniprot id: P49768)**: (a) Four contact maps constructed from coevolution analysis by DeepMetaPSICOV (DMP), RaptorX (RX), SPOT-Contact (SC), and the predicted structure by AlphaFold (AF) (only those contacts with probability >0.1 are shown). (b) W_3 scores for all residue positions based on the contact maps in (a), where red and blue dots mark residues with deleterious and neutral mutations, respectively. (c) Predicted structure by AlphaFold as colored by W_3 scores (red/blue for high/low values), where residues with deleterious and neutral mutations are shown as large and small balls, respectively (A79, V82, C92, V96, F105, L113, Y115, T116, P117, E120, N135, M139, I143, M146, T147, H163, W165, L166, S169, L171, L173, L174, G206, G209, I213, L219, A231, M233, L235, A246,

L250, A260, L262, C263, P264, G266, P267, R269, L271, R278, E280, L282, A285, L286, S289, D333, G378, G384, S390, L392, N405, A409, C410, A426, A431, P436, see https://www.uniprot.org/uniprotkb/ P49768/variant-viewer), and active-site residues are colored in green. (d) Four other network scores (MSF, δλ, C5 and C12) for all residue positions based on the contact maps in (a).

b(0,+)-type amino acid transporter 1 (Uniprot id: P82251): It forms a functional transporter complex that mediates the electrogenic exchange between cationic amino acids and neutral amino acids. Its dysfunction is linked to Cystinuria, an autosomal disorder characterized by impaired epithelial cell transport of cystine and dibasic amino acids in the proximal renal tubule and gastrointestinal tract (see https://www.uniprot.org/uniprotkb/P82251/entry#function).

The AlphaFold-predicted structure consists of a helical domain with deleterious mutations concentrating inside the core domain while those non-conserved residues are mostly located on the domain periphery (~~e.g.~~ N-terminal and C-terminal helices) (see Fig 3c). The active site consists of residues 43-47 and 233 and is also located in the core domain (colored green in Fig 3c). The contact maps predicted by three coevolution analysis tools are all similar to the contact map based on the AlphaFold structure (see Fig 3a) (except that RaptorX omitted some local contacts in residues 1-200). As a result, the network scores agree well between these methods (see Fig 3b). Reassuringly, those residues identified with high network scores are primarily within the central core and overlap with most deleterious mutations, while the peripheral regions feature low scores consistent with low sequence conservation. Among those top-10% hotspot residues (see Fig 3c), mutations I44T, A126T, and W230R were implicated in Cystinuria. *In vitro* measurements showed W230R has almost no transport activity, and it was

Formatted: Space After: 8 pt

proposed that W230 serves as a gate between two substrate-binding pockets and undergoes conformational changes to enable amino acid transport [58] . Although the A126T mutation is mildly dysfunctional [59], it is notable among a cluster of conserved residues with small sidechains in the contact regions of transmembrane helices, hinting for its possible role in helix-helix association and relative motions.

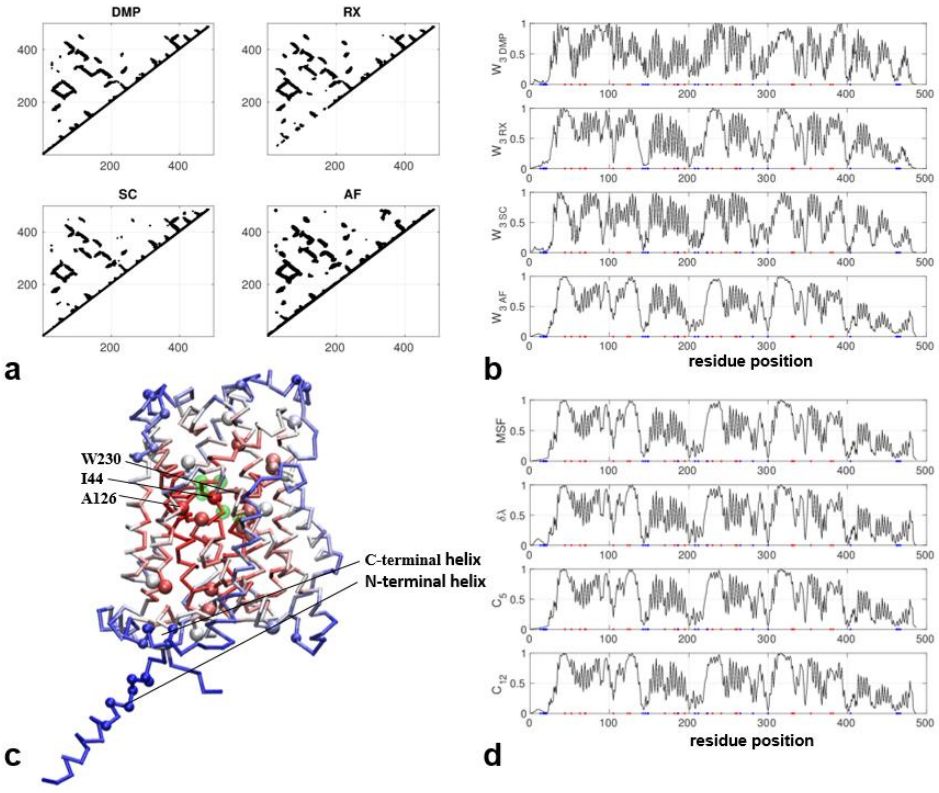


Figure 3. **Results for amino acid transporter 1 (Uniprot id: P82251)**: (a) Four contact maps constructed from coevolution analysis by DeepMetaPSICOV (DMP), RaptorX (RX), SPOT-Contact (SC), and the predicted structure by AlphaFold (AF) (only those contacts with probability >0.1 are shown). (b) W_3 scores for all residue positions based on the contact maps in

(a), where red and blue dots mark residues with deleterious and neutral mutations, respectively. (c) Predicted structure by AlphaFold as colored by W_3 scores (red/blue for high/low values), where residues with deleterious and neutral mutations are shown as large and small balls, respectively (V142,L223,I44,P52,G63,W69,A70,G105,T123,A126,V170,A182,I187,G195, A224,W230,I241,G259,P261,V330,A331,R333,A354,S379,A382, see https://www.uniprot.org/uniprotkb/ P82251/variant-viewer), and active-site residues are colored in green. (d) Four other network scores (MSF, δλ, C5 and C12) for all residue positions based on the contact maps in (a).

Lipoprotein lipase (Uniprot: P06858): As a key enzyme in triglyceride metabolism, it catalyzes the hydrolysis of triglycerides from circulating chylomicrons and very low density lipoproteins, thus playing an important role in lipid clearance from the blood stream, lipid utilization and storage (see https://www.uniprot.org/uniprotkb/P06858/entry#function).

The AlphaFold-predicted structure consists of an N-terminal helix, a central α/β domains, and a C-terminal β domain. Most deleterious mutations are concentrated inside the central domain while the non-conserved residues are mostly located on the periphery (including ~~e.g.~~ N-terminal helix and C-terminal domain) (see Fig 4c). The active site is comprised of a catalytic triad of S159, D183, and H268 [60] in the central domain (colored green in Fig 4c). The contact maps predicted by three coevolution analysis methods are similar to the contact map based on the AlphaFold structure (see Fig 4a). As a result, the network scores agree well between these methods (see Fig 4b) with minor differences in peripheral regions (~~e.g. in~~such as the N-terminal

Formatted

Formatted: Indent: First line: 0"

helix). As predicted, those residues identified with high network scores are primarily within the central domain overlapping with most deleterious mutations, while the peripheral N-terminal helix and C-terminal domain feature low scores consistent with low sequence conservation (see Fig 4c). Notably, some of them are found at the interface between the central domain and the C-terminal domain (circled in Fig 4c), possibly mediating inter-domain motions. Among those top-10% hotspot residues (see Fig 4c), T128, G132, H163, G169, G181, D183, P184, A185, D207, V208, H210, G222, V227, D231, I232, P234 and S271 are known to harbor pathogenic mutations in Hyperlipoproteinemia 1, an autosomal recessive metabolic disorder characterized by defective breakdown of dietary fats. Both H163 and G169 lie in helix 4 that constitutes part of the highly conserved beta-epsilon serine-alpha folding motif which is near S159 of the active site. Supporting their functional relevance, mutations H163R and G169E were found to abolish the enzymatic activity [61] . Near D183 (one of the catalytic triad), mutations G181S and P184R were found to abolish the catalytic activity [62] . Further from D183, conserved substations D207E and H210Q abolished the enzyme activity [63], and mutations D231E, I232S and P234L led to loss of the catalytic function [64] . These mutations may disrupt allosteric interactions with the central catalytic domain. Another conservative mutation S271T (near D183) also led to loss of enzyme activity [65]. Taken together, these residues may function by directly or indirectly coupling to the active site.

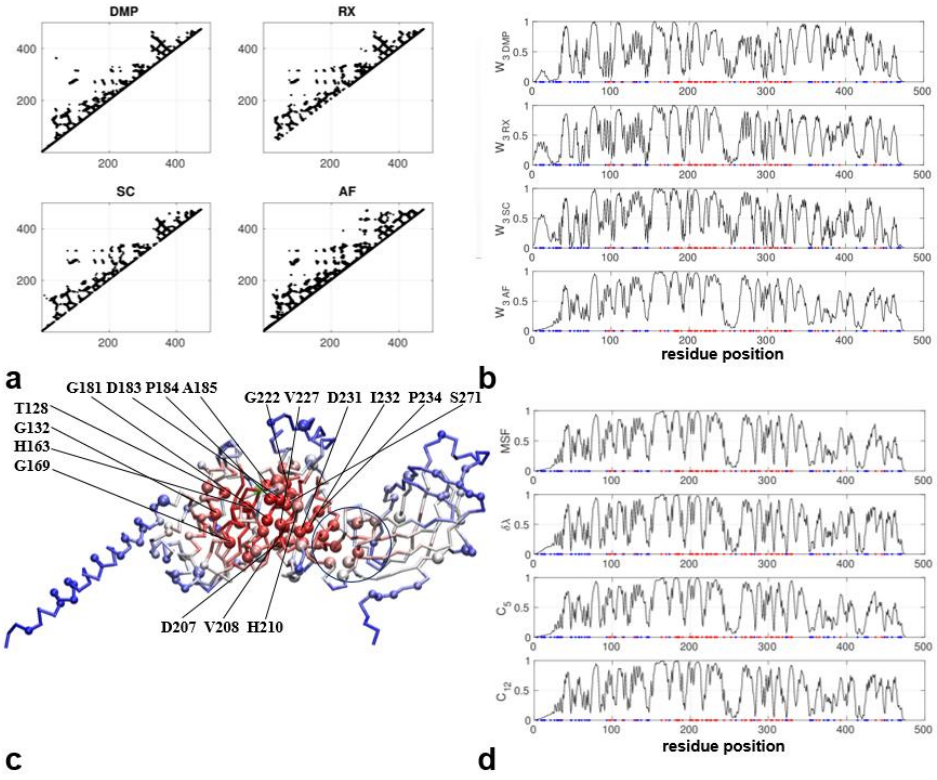


Figure 4. Results for Lipoprotein lipase (Uniprot id: P06858): (a) Four contact maps constructed from coevolution analysis by DeepMetaPSICOV (DMP), RaptorX (RX), SPOT-Contact (SC), and the predicted structure by AlphaFold (AF) (only those contacts with probability >0.1 are shown). (b) W_3 scores for all residue positions based on the contact maps in (a), where red and blue dots mark residues with deleterious and neutral mutations, respectively. (c) Predicted structure by AlphaFold as colored by W_3 scores (red/blue for high/low values), where residues with deleterious and neutral mutations are shown as large and small balls, respectively

(H71,A427,D36,N70,V96,A98,R102,W113,T128,G132,H163,G169,G181,D183,P184,A185,

G186,E190,S199,D201,A203,D207,V208,H210,G215, S220,I221,G222, K225,V227,D231, I232,P234,C243,I252,C266,R270,S271,D277,S278,L279,S286,Y289,F297,L303,C305, C310,L313,N318,S325,M328,L330,A361,S365,L392,E437,E437,C445,E448, see https://www.uniprot.org/uniprotkb/P06858/variant-viewer), and active-site residues are colored in green. (d) Four other network scores (MSF, δλ, C5 and C12) for all residue positions based on the contact maps in (a).

Conclusion

To conclude, we have combined machine learning, network analysis, and protein language models to develop a sequences-based variant site prediction method based on the protein residue contact networks which incorporate sequential, structural, dynamic, and interaction information simultaneously:

1. We build protein residue networks by exploiting three different state-of-the-art coevolution analysis tools (RaptorX, DeepMetaPSICOV, and SPOT-Contact) that complement each other.

2. We use three powerful machine learning algorithms (Random Forest, Gradient Boosting, and Extreme Gradient Boosting) to optimally combine 20 network centrality scores to accurately predict key residues as hot spots for disease mutations.

3. We train and validate our method using a dataset of 107 proteins rich in disease mutations, demonstrating its high accuracy in distinguishing between deleterious and neutral sites (with AUC of ROC ~ 0.84). Further improvement can be achieved after combining our method with the ESM-based method.

Formatted: Font color: Auto

Formatted: Font color: Auto

This study has established a useful strategy of combining an ensemble of network scores based on different coevolution analysis methods via machine learning to predict key variants sites of relevance to disease mutations. The code and dataset are made available to public to enable future developments and applications (see https://simtk.org/projects/hotspots).

For future work, it will be interesting to go beyond contact map predictions by integrating other scores derived from the co-evolution analysis (for example, see refs [66-68]) in our workflow, which may further boost the accuracy of variant site prediction.

References

~~reference~~

1. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, et al. (2021) Highly accurate protein structure prediction with AlphaFold. Nature 596: 583-589.
2. Baek M, DiMaio F, Anishchenko I, Dauparas J, Ovchinnikov S, et al. (2021) Accurate prediction of protein structures and interactions using a three-track neural network. Science 373: 871-876.
3. Terwilliger TC, Liebschner D, Croll TI, Williams CJ, McCoy AJ, et al. (2023) AlphaFold predictions are valuable hypotheses and accelerate but do not replace experimental structure determination. Nat Methods.
4. Al-Janabi A (2022) Has DeepMind's AlphaFold solved the protein folding problem? Biotechniques 72: 73-76.
5. Medina E, D RL, Sanabria H (2021) Unraveling protein's structural dynamics: from configurational dynamics to ensemble switching guides functional mesoscale assemblies. Curr Opin Struct Biol 66: 129-138.
6. Pak MA, Markhieva KA, Novikova MS, Petrov DS, Vorobyev IS, et al. (2023) Using AlphaFold to predict the impact of single mutations on protein stability and function. PLoS One 18: e0282689.
7. Steinegger M, Soding J (2018) Clustering huge protein sequence sets in linear time. Nat Commun 9: 2542.
8. Burley SK, Berman HM, Bhikadiya C, Bi C, Chen L, et al. (2019) RCSB Protein Data Bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. Nucleic Acids Res 47: D464-D474.
9. Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, et al. (2022) AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Res 50: D439-D444.
10. Bepler T, Berger B (2021) Learning the protein language: Evolution, structure, and function. Cell Syst 12: 654-669 e653.
11. Cheng J, Novati G, Pan J, Bycroft C, Zemgulyte A, et al. (2023) Accurate proteome-wide missense variant effect prediction with AlphaMissense. Science 381: eadg7492.
12. Hassan MS, Shaalan AA, Dessouky MI, Abdelnaiem AE, ElHefnawi M (2019) A review study: Computational techniques for expecting the impact of non-synonymous single nucleotide variants in human diseases. Gene 680: 20-33.
13. Niroula A, Urolagin S, Vihinen M (2015) PON-P2: prediction method for fast and reliable identification of harmful variants. PLoS One 10: e0117380.
14. Pejaver V, Urresti J, Lugo-Martinez J, Pagel KA, Lin GN, et al. (2020) Inferring the molecular and phenotypic impact of amino acid variants with MutPred2. Nat Commun 11: 5918.
15. Qi H, Zhang H, Zhao Y, Chen C, Long JJ, et al. (2021) MVP predicts the pathogenicity of missense variants by deep learning. Nat Commun 12: 510.
16. Yue P, Melamud E, Moult J (2006) SNPs3D: candidate gene and SNP selection for association studies. BMC Bioinformatics 7: 166.
17. Tang H, Thomas PD (2016) Tools for Predicting the Functional Impact of Nonsynonymous Genetic Variation. Genetics 203: 635-647.
18. Katsonis P, Koire A, Wilson SJ, Hsu TK, Lua RC, et al. (2014) Single nucleotide variations: biological impact and theoretical interpretation. Protein Sci 23: 1650-1666.

19. Singh A, Olowoyeye A, Baenziger PH, Dantzer J, Kann MG, et al. (2008) MutDB: update on development of tools for the biochemical analysis of genetic variation. Nucleic Acids Res 36: D815-819.
20. Bromberg Y, Rost B (2007) SNAP: predict effect of non-synonymous polymorphisms on function. Nucleic Acids Res 35: 3823-3835.
21. Ng PC, Henikoff S (2003) SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res 31: 3812-3814.
22. Pers TH, Timshel P, Hirschhorn JN (2015) SNPsnap: a Web-based tool for identification and annotation of matched SNPs. Bioinformatics 31: 418-420.
23. Zheng W, Tekpinar M (2009) Large-scale evaluation of dynamically important residues in proteins predicted by the perturbation analysis of a coarse-grained elastic model. BMC Struct Biol 9: 45.
24. Zheng W, Brooks BR, Doniach S, Thirumalai D (2005) Network of dynamically important residues in the open/closed transition in polymerases is strongly conserved. Structure 13: 565-577.
25. Ponzoni L, Bahar I (2018) Structural dynamics is a determinant of the functional significance of missense variants. Proc Natl Acad Sci U S A 115: 4164-4169.
26. Butler BM, Gerek ZN, Kumar S, Ozkan SB (2015) Conformational dynamics of nonsynonymous variants at protein interfaces reveals disease association. Proteins 83: 428-435.
27. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, et al. (2021) Applying and improving AlphaFold at CASP14. Proteins 89: 1711-1721.
28. Marks DS, Hopf TA, Sander C (2012) Protein structure prediction from sequence variation. Nat Biotechnol 30: 1072-1080.
29. Morcos F, Pagnani A, Lunt B, Bertolino A, Marks DS, et al. (2011) Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proc Natl Acad Sci U S A 108: E1293-1301.
30. Hopf TA, Scharfe CP, Rodrigues JP, Green AG, Kohlbacher O, et al. (2014) Sequence co-evolution gives 3D contacts and structures of protein complexes. Elife 3.
31. Marks DS, Colwell LJ, Sheridan R, Hopf TA, Pagnani A, et al. (2011) Protein 3D structure computed from evolutionary sequence variation. PLoS One 6: e28766.
32. Burger L, van Nimwegen E (2010) Disentangling direct from indirect co-evolution of residues in protein alignments. PLoS Comput Biol 6: e1000633.
33. Jones DT, Buchan DW, Cozzetto D, Pontil M (2012) PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. Bioinformatics 28: 184-190.
34. Halabi N, Rivoire O, Leibler S, Ranganathan R (2009) Protein sectors: evolutionary units of three-dimensional structure. Cell 138: 774-786.
35. Wang S, Sun S, Li Z, Zhang R, Xu J (2017) Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model. PLoS Comput Biol 13: e1005324.
36. Ma J, Wang S, Wang Z, Xu J (2015) Protein contact prediction by integrating joint evolutionary coupling analysis and supervised learning. Bioinformatics 31: 3506-3513.
37. Kandathil SM, Greener JG, Jones DT (2019) Prediction of interresidue contacts with DeepMetaPSICOV in CASP13. Proteins 87: 1092-1099.
38. Jones DT, Singh T, Kosciolek T, Tetchner S (2015) MetaPSICOV: combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins. Bioinformatics 31: 999-1006.
39. Jones DT, Kandathil SM (2018) High precision in protein contact prediction using fully convolutional neural networks and minimal sequence features. Bioinformatics 34: 3308-3315.

40. Hanson J, Paliwal K, Litfin T, Yang Y, Zhou Y (2018) Accurate prediction of protein contact maps by coupling residual two-dimensional bidirectional long short-term memory with convolutional neural networks. Bioinformatics 34: 4039-4045.
41. Yan W, Yu C, Chen J, Zhou J, Shen B (2020) ANCA: A Web Server for Amino Acid Networks Construction and Analysis. Front Mol Biosci 7: 582702.
42. Amitai G, Shemesh A, Sitbon E, Shklar M, Netanely D, et al. (2004) Network analysis of protein structures identifies functional residues. J Mol Biol 344: 1135-1146.
43. Velickovic P (2023) Everything is connected: Graph neural networks. Curr Opin Struct Biol 79: 102538.
44. Butler BM, Kazan IC, Kumar A, Ozkan SB (2018) Coevolving residues inform protein dynamics profiles and disease susceptibility of nSNVs. PLoS Comput Biol 14: e1006626.
45. Chasman D, Adams RM (2001) Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: structure-based assessment of amino acid variation. J Mol Biol 307: 683-706.
46. Gerek ZN, Ozkan SB (2011) Change in allosteric network affects binding affinities of PDZ domains: analysis through perturbation response scanning. PLoS Comput Biol 7: e1002154.
47. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, et al. (2010) A method and server for predicting damaging missense mutations. Nat Methods 7: 248-249.
48. Meier J, Rao R, Verkuil R, Liu J, Sercu T, et al. (2021) Language models enable zero-shot prediction of the effects of mutations on protein function. bioRxiv: 2021.2007.2009.450648.
49. Vihinen M (2020) Problems in variation interpretation guidelines and in their implementation in computational tools. Mol Genet Genomic Med 8: e1206.
50. Ashkenazy H, Erez E, Martz E, Pupko T, Ben-Tal N (2010) ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. Nucleic Acids Res 38: W529-533.
51. Zheng W, Brooks BR, Thirumalai D (2006) Low-frequency normal modes that describe allosteric transitions in biological nanomachines are robust to sequence variations. Proc Natl Acad Sci U S A 103: 7664-7669.
52. Zheng W (2016) Probing the structural dynamics of the SNARE recycling machine based on coarse-grained modeling. Proteins.
53. Delgado J, Radusky LG, Cianferoni D, Serrano L (2019) FoldX 5.0: working with RNA, small molecules and a new graphical interface. Bioinformatics 35: 4168-4169.
54. Veiga-da-Cunha M, Gerin I, Chen YT, de Barsy T, de Lonlay P, et al. (1998) A gene on chromosome 11q23 coding for a putative glucose- 6-phosphate translocase is mutated in glycogen-storage disease types Ib and Ic. Am J Hum Genet 63: 976-983.
55. Hiraiwa H, Pan CJ, Lin B, Moses SW, Chou JY (1999) Inactivation of the glucose 6-phosphate transporter causes glycogen storage disease type 1b. J Biol Chem 274: 5532-5536.
56. Wolfe MS, Xia W, Ostaszewski BL, Diehl TS, Kimberly WT, et al. (1999) Two transmembrane aspartates in presenilin-1 required for presenilin endoproteolysis and gamma-secretase activity. Nature 398: 513-517.
57. Sun L, Zhou R, Yang G, Shi Y (2017) Analysis of 138 pathogenic mutations in presenilin-1 on the in vitro production of Abeta42 and Abeta40 peptides by gamma-secretase. Proc Natl Acad Sci U S A 114: E476-E485.
58. Yan R, Li Y, Shi Y, Zhou J, Lei J, et al. (2020) Cryo-EM structure of the human heteromeric amino acid transporter b(0,+)AT-rBAT. Sci Adv 6: eaay6379.
59. Font MA, Feliubadalo L, Estivill X, Nunes V, Golomb E, et al. (2001) Functional analysis of mutations in SLC7A9, and genotype-phenotype correlation in non-Type I cystinuria. Hum Mol Genet 10: 305-316.

60. Emmerich J, Beg OU, Peterson J, Previato L, Brunzell JD, et al. (1992) Human lipoprotein lipase. Analysis of the catalytic triad by site-directed mutagenesis of Ser-132, Asp-156, and His-241. J Biol Chem 267: 4161-4165.
61. Reina M, Brunzell JD, Deeb SS (1992) Molecular basis of familial chylomicronemia: mutations in the lipoprotein lipase and apolipoprotein C-II genes. J Lipid Res 33: 1823-1832.
62. Bruin T, Tuzgol S, van Diermen DE, Hoogerbrugge-van der Linden N, Brunzell JD, et al. (1993) Recurrent pancreatitis and chylomicronemia in an extended Dutch kindred is caused by a Gly154-->Ser substitution in lipoprotein lipase. J Lipid Res 34: 2109-2119.
63. Haubenwallner S, Horl G, Shachter NS, Presta E, Fried SK, et al. (1993) A novel missense mutation in the gene for lipoprotein lipase resulting in a highly conservative amino acid substitution (Asp180-->Glu) causes familial chylomicronemia (type I hyperlipoproteinemia). Genomics 18: 392-396.
64. Gotoda T, Yamada N, Kawamura M, Kozaki K, Mori N, et al. (1991) Heterogeneous mutations in the human lipoprotein lipase gene in patients with familial lipoprotein lipase deficiency. J Clin Invest 88: 1856-1864.
65. Hata A, Emi M, Luc G, Basdevant A, Gambert P, et al. (1990) Compound heterozygote for lipoprotein lipase deficiency: Ser----Thr244 and transition in 3' splice site of intron 2 (AG----AA) in the lipoprotein lipase gene. Am J Hum Genet 47: 721-726.
66. Bisardi M, Rodriguez-Rivas J, Zamponi F, Weigt M (2022) Modeling Sequence-Space Exploration and Emergence of Epistatic Signals in Protein Evolution. Mol Biol Evol 39.
67. Cocco S, Feinauer C, Figliuzzi M, Monasson R, Weigt M (2018) Inverse statistical physics of protein sequences: a key issues review. Rep Prog Phys 81: 032601.
68. Rodriguez-Rivas J, Croce G, Muscat M, Weigt M (2022) Epistatic models predict mutable sites in SARS-CoV-2 proteins and epitopes. Proc Natl Acad Sci U S A 119.

Supporting Information

S1 Table. Evaluation of 20 network scores based on protein residue contact maps constructed from 3 coevolution analysis tools (DeepMetaPSICOV, RaptorX, and SPOT-Contact)

Score	AUC* of DeepMetaPSICOV	AUC* of RaptorX	AUC* of SPOT-Contact
C1	0.75±0.13	0.78±0.14	0.75±0.15
C2	0.75±0.17	0.75±0.19	0.76±0.19
C3	0.78±0.10	0.75±0.15	0.70±0.15
C4	0.65±0.12	0.52±0.15	0.61±0.07
C5	0.78±0.17	0.78±0.17	0.78±0.18
C6	0.63±0.16	0.56±0.19	0.65±0.17
C7	0.77±0.11	0.61±0.20	0.72±0.16
C8	0.65±0.12	0.52±0.15	0.61±0.07
C9	0.79±0.13	0.78±0.16	0.75±0.16
C10	0.76±0.12	0.75±0.14	0.69±0.15
C11	0.80±0.13	0.78±0.17	0.78±0.17
C12	0.78±0.16	0.80±0.14	0.80±0.17
C13	0.75±0.17	0.74±0.18	0.76±0.19
$\delta\lambda$	0.80±0.13	0.78±0.14	0.78±0.16
MSF	0.81±0.14	0.79±0.15	0.80±0.17
W_1	0.81±0.14	0.79±0.15	0.80±0.17
W_2	0.81±0.15	0.77±0.15	0.79±0.17
W_3	0.81±0.14	0.78±0.15	0.80±0.16
W_∞	0.80±0.13	0.77±0.14	0.78±0.16
W_s	0.80±0.13	0.78±0.13	0.78±0.16

* mean ± standard-deviation

Formatted: Font: 18 pt

Formatted Table

Formatted: Subscript

Response to Editor

I have addressed the following additional journal requirements:

1. I have reformatted the manuscript to comply with PLOS ONE's style requirements (journals.plos.org/plosone/s/submission-guidelines), including those for file naming.

2. No human participants are involved in this study.

3. To comply with PLOS ONE guidelines on code sharing I will make the code available at https://simtk.org/projects/hotspots .

4. Since I use WORD to write the manuscript, the PLOS LaTeX template is not applicable.

5. I have corrected the grant information so it matches between the 'Funding Information' and 'Financial Disclosure' sections. I also provide the correct grant numbers (3R01NS108750) for the awards in the 'Funding Information' section.

6. In the financial disclosure "This study is funded by a grant from NIH…", I would like to add this amended Role of Funder statement: "The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript."

7. I have made all relevant data available without restriction at https://simtk.org/projects/hotspots .

I have revised the paper to address the following additional Editor Comments: In particular, the abstract and introduction should be rewritten based on the provided comments. In addition, the proposed methodology should be better described and justified. Finally results should be better presented as suggested by reviewers.

Response to Reviewers

Response to Reviewer #1

We thank reviewer 1 for the constructive comments and suggestions!

Abstract

- The abstract in this paper did not adequately capture the details expected in an abstract. Background to the research was not introduced. Problem being addressed was not efficiently stated. Previous methods that have tried addressing the problems were not highlighted. For instance, the author stated that "To meet this challenge, we build upon recent progress in machine learning, network analysis, and protein language models, and develop a …" without actually highlighting the previous work done. Majority of the content in the abstract are on the authors' finding. The author is expected to provide a basic or simple background to the research. Followed by a brief description of what has been previously done before stating what the problems are from what has been done. The author can be more specific on the keywords which shouldn't be more than 6.

We have revised Abstract to include more background information (including problem to address, previous methods and their limitations). We have reduced the number of keywords to no more than six.

Introduction

- The authors should minimize the use of 'e.g' both in the introduction section and the abstract.

We have removed most 'e.g' in the paper.

- Citation in the body of the introduction is quite few. More citations and references should be provided.

We have added more references in Introduction (with total 48 references cited).

- Kindly provide justification for this claim "…it is now feasible to predict static structures for many proteins of interest given their sequences".

We have added new references to substantiate the above claim with some caveats (line 42).

- It appears that the last paragraph of the introduction is exactly same as some major part of the abstract word for word. Authors are encouraged to avoid self-plagiarism.

We have rewritten this paragraph and Abstract to avoid duplications between them.

- Similar to the abstract, information in the introductory part of this paper is insufficient.

We have tried our best to give a detailed introduction to the background and literature relevant to this study (with total 48 references cited). We will appreciate it if the reviewer could kindly give more specific comments if any information is still missing.

- Also, authors are advised to provide a brief description of the different sections of the manuscript at the end of the introduction.

We have added a brief outline of the Methods and Results sections at the end of Introduction (line 125).

Materials and Methods

•The subsequent section after the introduction should be captioned “Materials and Methods” as opposed to the caption used by the author.
We have renamed the Method section to “Materials and Methods”.

•Each section and subsection should be numbered accordingly. Currently, sections and subsections are not identifiable.
We have numbered all subsections of “Materials and Methods” in the order of the proposed workflow.

•Furthermore, for easier flow and understanding of the methodology, a framework or an algorithm of the methodology could be added. This would provide readers with a conceptualized view of the methodology.
We have added a summary of the workflow of our method at the beginning of “Materials and Methods” (line 131).

•The uniport Id of the 107 protein sequences collected can be provided as a supplementary file.
Detailed information about the dataset of 107 protein sequences and variants is available at the following site: https://simtk.org/projects/hotspots

•The author claimed to have used Random Forest, Gradient Boosting Classifier and Extreme Gradient Boosting Classifier. Kindly provide a justification for the choice of this machine learning methods
These three methods were chosen because they have performed successfully in machine learning contests in Kaggle. They are also relatively cheap to train and optimize compared with the deep learning methods (see line 276).

•In addition, what parameters were tuned by the author? Did the author used Optuna for all the ML methods. How is the result prior to the use of hyper-parameter tuning technique?
We added details of hyper-parameters at lines 266, 270, 274. Yes, we used Optuna for each of the three ML methods. While the resulting improvement is modest (relative to the default parameters), it is a common practice in ML to perform task-specific hyper-parameters tuning. We have run Optuna multiple times to ensure the resulting best metric is reproducible.

•Why was AUC of ROC used as the metric for assessing prediction quality?
The AUC is a standard metric for evaluating binary classifiers based on the ROC curve of sensitivity and specificity. The ROC curves are also used in other computational papers for variant predictions (see line 162).

Results and Discussion

•Most importantly, results obtained by the author should be presented while being discussed instead of being added to the supplementary or being placed at the end of the paper. This makes it difficult to understand the result being discussed. Some of the figures should also be presented as they are being discussed.

We have moved figures and tables to where they were discussed.

Conclusion

•The conclusion section is completely missing in this paper. Author is encourage to provide a conclusion section alongside a summarized and itemized key findings from the research.
We have added Conclusion to summarize our key findings.

Response to Reviewer #2

We thank reviewer 2 for the constructive comments and suggestions!

Major points

• Sequences-based coevolution methods return a prediction of the protein structure from multiple sequence alignments (MSAs). It is not clear to me how these MSAs (or a unique MSA?) are designed. The authors sometimes refer to them (or to it), and sometimes they put the focus on the 107 sequences used in the training and testing of the binary classifier but it is not clear to me how they are chosen, and if they belong to the initial MSA or not. I have appreciated the final focus on four sequences, even though the performances, and a comparison to other techniques, of the new pipeline (for instance the AUC) should be included in the discussion.
To clarify, the coevolution-based methods that we used (RaptorX, DeepMetaPSICOV, and SPOT-Contact) take a protein sequence as input and then predict a residue-residue contact probability matrix (not the protein structure) based on MSA. Our method does not utilize the MSA directly. The 107 protein sequences were chosen from the HumVar database (see line 143), and they do not belong to a specific MSA. Instead, each of these sequences was used to build its own MSA by the above coevolution-based methods.

We have compared our sequences-based method with alternative methods based on structures (AlphaFold) or physical force fields (FoldX) or protein language model (ESM) using a testing set of 28 proteins (see Tables 2 and 4). We also added additional comparison with PolyPhen-2 (see line 397).

• Linked to the first question, if the analysis is performed on a unique dataset, i.e. a unique MSA of a protein family, I encourage the authors to repeat these experiments on different protein domains, as a stress test to their new pipeline.
To clarify, our evaluation is not performed on a unique MSA of a particular protein family. Instead, it is based on a diverse dataset of 107 proteins from various different protein families, which contain 97 dissimilar proteins (with their pairwise sequence identity < 30%) (line 147).

• In the third step of the workflow, the authors split the 107 protein sequences into a train and a test set. How is this subdivision decided? One should in principle check whether the sequences in the two sets are sufficiently "distant" (for instance using clustering analysis) or repeat the procedure for different assignments into the two subsets. Additionally, how does the choice of the size of these two sets affect the results?
The train/test split is purely random: 79 for training and 28 for testing. We have checked the sequence similarity between the two sets, only 1 training-set protein has sequence identity >30% with proteins of the testing set. In fact, 97/107 proteins have pairwise sequence identity <30%. We repeated training and testing on this reduced dataset of 97 dissimilar proteins, and the resulting AUCs are only slightly higher (by <0.03), suggesting that the few similar sequences did not markedly change the training/testing results.
The choice of train/test split ratio (~75/25) is based on common practice of ML (see https://onlinelibrary.wiley.com/doi/full/10.1002/sam.11583). We also tried other sizes of testing set (18 and 38) which only slightly increase the resulting AUCs by 0.01~0.04. So the training/testing results are not sensitive to this choice.

• Coevolution methods, together with a proxy for pairwise contact prediction, allow for an estimate of the degree of deleteriousness of point and pairwise mutations. In recent developments (see Refs. 37-42 in [1] and also more recent works in [2-3]), the authors show that the sequence "energy" of the coevolution models can be interpreted as (negative) protein fitness which, indeed, correlates well with deep mutational scanning-based measures. Since this coevolution information is exploited in the first step of the presented pipeline, I would like to ask the following questions:

– Can the authors display how well these methods alone perform compared to the author's workflow?
Although it would be useful to use ΔE_DCA defined in ref[3] to predict variant effects, such calculation is not supported by the co-evolution-based contact prediction methods used here (RaptorX, DeepMetaPSICOV, and SPOT-Contact), and is therefore beyond the scope of this study. This study is limited to the use of residue contact maps as predicted by co-evolution analysis, rather than fully exploiting all scores derived from the co-evolution analysis of MSA.

– Can this information be integrated (together with the pure network scores) in the second and third steps?
Yes, we expect potentially fruitful integration of our network-based methods with other co-evolution-based scores like ΔE_DCA to improve the prediction of variant effects. We have mentioned this and cited ref [1-3] in the discussion of future work (see line 590).

• It is not clear to me why the twenty measures used in the second step (if I have understood correctly, they are associated with network centrality properties) are the correct (or the most informative) metrics to cope with sequence hotspots. I would expect that some important hotspots may be related to the interaction with other proteins, and, therefore, they may be "far" from the core of the folded protein network. Also, other information like the electrical-chemical properties of the amino acids is neglected. Can the authors comment on them?
We agree that not all hotspot residues can be predicted by our method which focused on intra-protein residue contacts, and some hotspot residues may not possess high centrality scores (as mentioned above by the reviewer). However, we have shown that our method is competitive with alternative methods (ESM and PolyPhen-2), supporting the value of centrality scores as informative predictors for disease mutations. Further, our method provides new predictive features that complement other scores (such as ΔE_DCA), and together they may enable more accurate predictions of variant effects (see line 590).

• Pag. 15. The evaluation of each single score (among the twenty?) against the different structures is not well described. The comparison is made according to which measure? The final binary classifier? If this is the case, does it mean that the authors compared the importance of each score independently and then all together?
For the evaluation of each single score, we sort all testing-set variants by that score and predict a variant deleterious/neutral if its score is above/below a cutoff value. This resulted in an ROC curve from which we calculated AUC (line 302).
After assessing the individual scores independently, we then used ML to combine them to train a binary classifier which was then assessed with the AUC of ROC (see Table 4).

• Overall, the presentation of the results is a little confusing to me. Comparisons seem to be made "internally" by changing the metrics in the second step, of the protein structure prediction in the first step. On page 19, the authors mention a comparison between their algorithm and Seq-GNM,

PolyPhen-2 but they run on different datasets, so the final performances may not be compared. I believe that a more "fair" comparison with other state-of-the-art techniques should be presented.
We agree that fair comparisons with alternative methods like PolyPhen-2 are desirable. However, since our training/testing datasets were taken from HumVar, and PolyPhen-2 has been trained on this dataset, it is not possible to compare their performance objectively. With this caveat in mind, we have evaluated PolyPhen-2 based on the same 28 testing-set proteins and their variants, and found the AUC of PolyPhen-2 to be 0.85 (see line 397), which is close to our method (after combining 3 co-evolution analyses, see Table 4). Additionally, we have compared our method with another state-of-the-art method based on ESM, which gave AUC~0.83. Therefore, our method is competitive with these alternative methods. To be clear, our main goal is to complement rather than compete with existing methods. Indeed, we have shown that the combination of our method with ESM has yielded better performance than ESM alone (see Table 4).

• The ROC and AUC metrics are presented as final comparison metrics. Is the value presented in the manuscript an average value among the test sequences? If this is the case, what about the standard error associated with it?
To clarify, the AUCs in Table 2 are calculated from the ROC for all variants of the 28 testing set proteins. We also calculated AUCs for each protein alone, and their means and S.D. are shown in Table S1. The cumulative AUCs in Table 2 are comparable to the mean AUCs of individual proteins in Table S1.

Minor points

• Would it be possible to swap the Methods section and the Results section? Also, some parts are frequently repeated in both Methods and Results.
I am afraid not because the PLOSONE format requires the Methods section precedes the Results section.
We have reduced methodological details in Results to avoid repetitions.

• Within the Methods section, in my opinion, the paper would gain readability if the sections followed the main pipeline of the method (now the third step is described before the first one).
We have numbered the subsections in the order of the workflow/pipeline (starting with the datasets). The workflow is summarized at line 131.

References

[1] Inverse statistical physics of protein sequences: a key issues review S Cocco, C Feinauer, M Figliuzzi, R Monasson, M Weigt Reports on Progress in Physics 81 (3), 032601

[2] Modeling sequence-space exploration and emergence of epistatic signals in protein evolution M Bisardi, J Rodriguez-Rivas, F Zamponi, M Weigt Molecular biology and evolution 39 (1), msab321

[3] Epistatic models predict mutable sites in SARS-CoV-2 proteins and epitopes J Rodriguez-Rivas, G Croce, M Muscat, M Weigt Proceedings of the National Academy of Sciences 119 (4), e2113118119

Data Review URL