Structure-Based Pathogenicity Relationship Identifier (SPRI): A Novel Structure-Guided Method to Evaluate Pathological Effect of Missense Mutations

Boshen Wang [1], Xue Lei [1], Wei Tian [1], Alan Perez-Rathke [1], Yan-Yuan Tseng [2], Jie Liang [1,* *1]

[1]Center for Bioinformatics and Quantitative Biology, Richard and Loan Hill Department of Biomedical Engineering, University of Illinois at Chicago, 820 S. Wood Street, Chicago, IL 60612, USA
[2] Center for Molecular Medicine and Genetics, Biochemistry and Molecular Biology Department, School of Medicine, Wayne State University, 540 E. Canfield, Scott Hall 3220, Detroit, MI 48201

Abstract

Missense mutations are among the most frequently occurring variants in exon regions and may lead to pathogenic phenotypes. Computational methods predicting effects of missense mutations play important roles in assessing pathogenicity of these variants. While a number of methods have been developed based on analysis of sequence conservation, co-evolution, and protein structures, effectively determining mutation effects on biochemical function remains challenging. Here we report the method of Structure-Based Pathogenicity Relationship Identifier (SPRI) that can accurately evaluate pathological effects of missense mutations and identify those that are deleterious. In addition to sequence analysis, our method quantifies short-, intermediate-, and long-range topological interactions

*To whom correspondence should be addressed. Email: jliang@uic.edu

at atomic and residue-level through geometric computation. Our method also explicitly computes surface pockets, and considers mutagenic changes in biophysical properties. The SPRI method performs favorably in identifying deleterious mutations and in quantifying their pathogenicity compared to current state-of-the-art methods in Mendelian diseases, as measured by several performance metrics. In addition, SPRI captures common properties among pathological missense mutations of both germline and somatic origins. Furthermore, the pathogenicity model is transferable across Mendelian diseases and cancer types as SPRI makes accurate predictions on effects of cancer driver mutations.

1 Introduction

Whole-genome sequencing (WGS) and whole-exome sequencing (WES) provide powerful means to assess genetic diversity among individuals [1, 2]. On average, an individual has hundreds of variants in the coding regions [3, 4, 5]. Among these variants, missense mutations have the highest frequency of occurrence and can potentially affect molecular functions upon residue substitutions [5, 6, 7, 8]. Therefore, it is important to determine whether a missense mutation leads to neutral or deleterious phenotypic changes.

A number of methods have been developed to evaluate potential pathological effects of missense mutations and identify deleterious mutations, including EVmutation, FATHMM, LIST, PMut, Polyphen-2, PROVEAN, and SIFT [10, 11, 12, 14, 13, 15, 9]. Based on sequence alignment and analysis of evolutionary conservation [10, 11, 12, 14, 13, 15, 9], these methods integrate prior knowledge of protein functions [11], inferred physico-chemical properties [14], evolutionary distances by taxonomy [12], and sequence co-evolution [10] to identify deleterious missense mutations. They have been widely used to study mutation effects. However, there are limitations to these methods. Some were trained

on particular datasets and exhibit deteriorated performance on other datasets, indicating lack of transferability [12, 11, 17]. Others require annotated knowledge of functional sites and functional domains, resulting in diminished prediction coverage as annotation is not uniformly available for each mutation [10] (see Table 1).

Complementing the extraordinary progress in sequencing, the structures of a large number of proteins (over 167,000 as of June, 2021) have been resolved by X-ray, NMR and Cryo-EM techniques [18, 21, 20, 19]. These structures provide rich information on how proteins carry out their functions. Several methods, including topoSNP, Cancer3D, Mutation3D, and Hotspot3D, were developed to map the spatial positions of mutation sites, to identify clusters of mutations, or to determine whether mutations are close to known functional sites or domains [26, 22, 23, 24]. In addition, the methods of Rhapsody and DAMpred were developed that utilize structural information to predict pathological effects of missense mutations and to identify those that are deleterious [16, 28, 27]. Rhapsody considers the dynamic behavior of proteins based on elastic network models constructed from corresponding protein structures [29]. The inferred dynamic properties are then integrated with Polyphen-2-type

of sequence analysis for predictions [16, 28]. To addresses the problem that many proteins lack experimentally determined structures, DAMPRED incorporates homology-modelled protein structures and pharmacophore features, and employs a Bayesian-guided neutral network to classify mutations [27]. These methods show that the incorporation of protein structures leads to improved characterization of the biochemical properties of the mutation sites.

However, it is challenging to determine the relevant factors affecting mutation effects from the myriad of information contained in proteins structures. Biochemical functions require specific spatial arrangement of residues and atoms to present a binding or functional surface with required biochemical properties [30], such as the general properties of electrostatic environment necessary for electron transfer in enzyme reactions [31]. Mutations may change properties of the binding surfaces or alter important atomic interactions near binding surfaces, which may impact the microenvironment of biochemical reactions. These mutations are likely to result in altered biochemical reactions and hence may be deleterious. Such information is encoded in structural properties of proteins, *e.g.*, atomic interactions, residue effects on the surrounding environment, the tertiary arrangements of functionally important residues, as well as solvent accessibility of the surface. These structural features complement evolutionary signals extracted from sequences. However, current structure-based methods do not yet consider detailed properties of the binding or functional surfaces of proteins, where mutational effects may be strong.

Here we present a new method called SPRI (**S**tructure-Based **P**athogenicity **R**elationship **I**dentifier), which provides quantitative assessment of pathological

effects of missense mutations and identifies deleterious mutations. Our method is based on in-depth structure analysis, including novel interaction profiles of short-range atomic interactions, intermediate-, and long-range residue interaction derived from atomic interactions, all are based on computation of the alpha-shapes of protein structures [33, 32]. We further investigate structures of surface pockets [32, 34, 26, 35, 36], solvent accessible surface area [37], and salt-bridge interactions of the proteins. Changes in biophysical properties of the side-chains between the wild-type and the substituted residues are also considered. The structural information is then integrated with evolutionary signals of the mutation site derived from multiple sequence alignment. A random forest predictor is constructed using this information as input for prediction [38].

We have compared our method with several state-of-the-art methods, including EVmutation, FATHMM, LIST, PMut, Polyphen-2, PROVEAN and Rhapsody [10, 11, 12, 14, 13, 15, 28]. Using benchmark data sets with reduced inherent bias of uneven representations of neutral and deleterious mutations, we have examined how well each method performs in predicting deleterious germline mutations of Mendelian diseases. Overall, our methods outperforms other methods on most quality metrics, including Area Under the Receiver Operating Curve (AUROC), Area Under the Precision-Recall Curve (AUPRC), Matthews correlation coefficient (MCC), as well as F-1 score. In addition, our method is highly transferable and can effectively evaluate pathogenicity of mutations regardless of disease types. Among somatic mutations collected from cancer samples, our method exhibits strong sensitivity in identifying confirmed cancer driver mutations, without the need of re-training using different datasets or requiring additional threshold adjustment. In contrast, current methods are inconsistent in such predictions, with higher rate of false negative predictions [11, 28].

Overall, our method allows detection of deleterious mutations contributing to single gene disorders such as Mendelian-type diseases, as well as complex-trait diseases such as cancer. Our results demonstrate that accurate assessment of pathological effects of missense mutation can be obtained when structural information and biophysical constraints are properly extracted and integrated with evolutionary information. We expect that our method can be broadly applied to assess missense mutations regardless of the mutational origins (germline or somatic), or disease types.

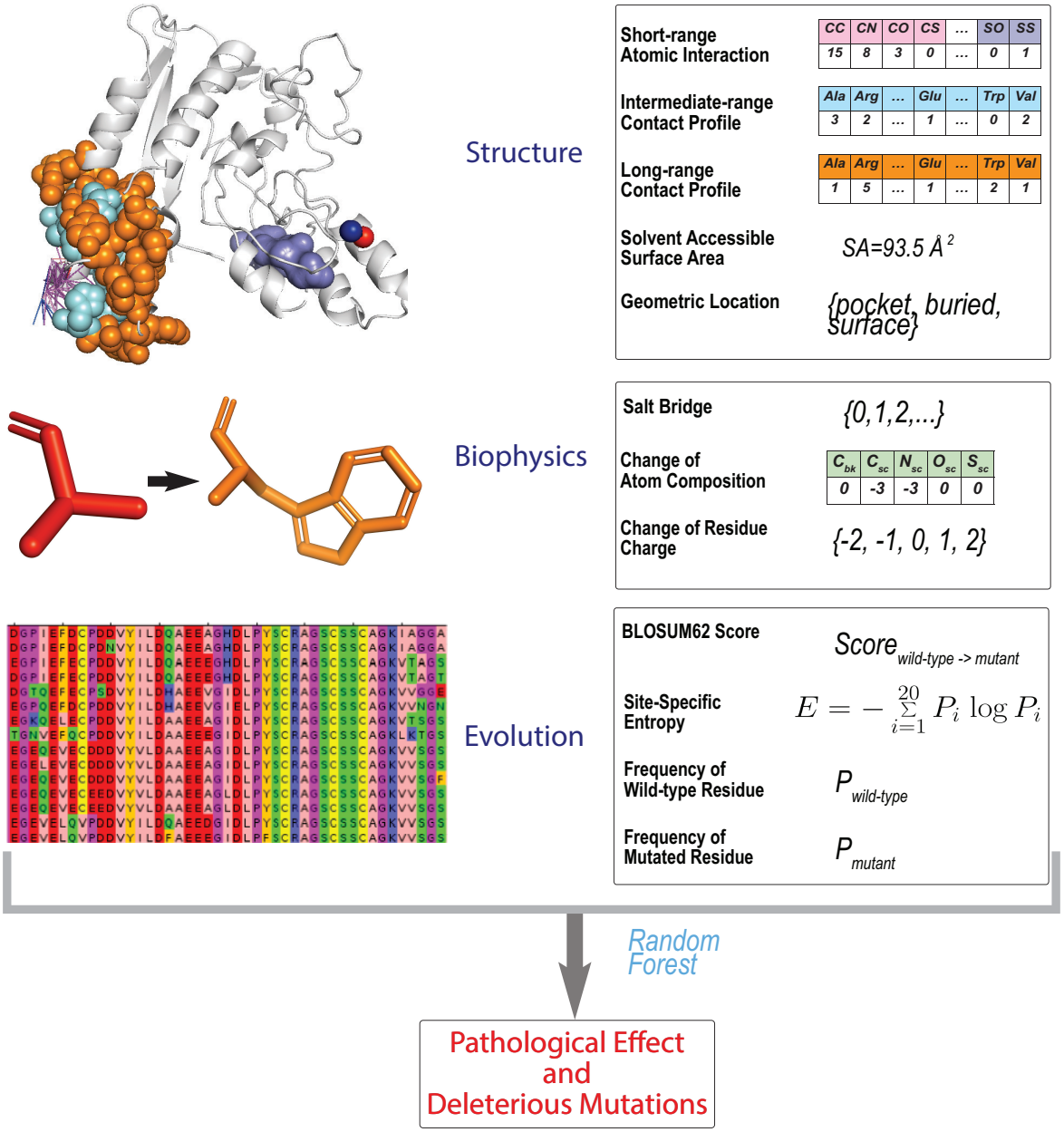

Figure 1: Overview of the SPRI method. SPRI utilizes three types of features, namely, protein structures, biophysical properties, and evolution signals. Structural features include the novel atomic and residue interaction profiles computed using alpha shapes, which capture the physical chemical micro-environment of the mutation site. In addition, category of geometric shape at the mutation site and at neighboring residues, as well as solvent accessible surface area are incorporated. The biophysical properties includes salt bridges formed by ionizable residues, changes in charge upon substitutions, as well as changes in side-chain and backbone atoms. A knowledge database containing ∼24 millions protein sequences from reference proteomes of eukaryotic species in UniProt is also constructed to identify homologs, as multiple sequence alignment can be constructed to obtain site-specific metrics. These features, except geometric location, are vectorized into numerical values. A random forest classifier is then trained to generate probability measure of the likelihood of pathological effect, from which a binary classification is made.

2 MATERIALS AND METHODS

The overall architecture of our novel method is presented in figure 1.

2.1 Construction of Datasets

We obtain the HumDiv, HumVar datasets from the Polyphen-2 website `http://genetics.bwh.harvard.edu/pph2/`, and the PredictSNP dataset from `https://loschmidt.chemi.muni.cz/predictsnp/`. We map all mutations to the same version of the canonical protein sequence of Uniprot Jan 2021 release [39]. This reconciles occasional sequence inconsistency among different reference sequences. The human protein sequence and sequence database of eukaryotic reference proteome is accessed from `https://ftp.uniprot.org`.

We then discard redundant mutations, remove conflicting mutations with both neutral and deleterious labels for the same mutation. Furthermore, we select only those proteins that contain both deleterious and neutral missense mutations. To retrieve the experimental determined Pdb structures, we employ the SeqMapPDB tool to obtain full-coverage structures and partial structural domains, respectively [18, 40]. We then generate the UnifyPDBFull and UnifyPDBAcceptable benchmark datasets. The former contains proteins whose full structures are known, and the latter contains proteins with both full-coverage structures and domains of structures with partial-coverage . The UnifyPDBFull dataset contains 4,231 deleterious variants and 2,791 neutral variants, which are derived from 252 proteins and are mapped to 252 polypeptide chains in PDB structures. The UnifyPDBAcceptable dataset contains 5,999 deleterious variants and 3,485 neutral variants, which are derived from 377 proteins, and are mapped to 444 Pdb structural chains.

2.2 Structure-Derived Features

The structural coordinates of proteins are retrieved from the PDB database at `https://ftp.rcsb.org/pub/pdb/`. We calculate the weighted alpha-shape using the van der Waals (VDW) radius of each chemical element, with water molecule probe of 1.4 Å radius. Atomic interactions are defined by alpha shapes, which captures exact near-neighboring atomic contacts that are dual to Voronoi boundaries separating two inter-residue atoms [41]. We then construct the inter-residue interaction profiles. For short-range atomic interactions, we define 16 types of element pairs, and vectorize spatial contact information into integer values for each interaction type. We use the Breadth-First Search (BFS) algorithm to obtain atom-interactions at residue level in intermediate- and long-distance range [42]. These are converted to 20-element vectors, where each element represents the number of each amino acid type occurs at the assigned distance range. We use the CASTP server to compute the solvent accessible surface area (SA) and assign property of geometric location for each residue [43]. A salt bridge is considered to exist if the distance between the oxygen atoms in an acidic residue and the nitrogen atoms in a basic residue is within 3.2 Å [44].

2.3 Encoding Biophysical Changes Introduced by Variants

We regard Asp and Glu as negatively (-1) ionizable residues. Arg, His and Lys as positively (+1) ionizable residues. All remaining amino acids are labelled as neutrally charged (0). For every mutation pattern, we calculate the change of charges by taking the difference of the values of charge labels, which are integer values in the range of -2 to 2.

Changes in atomic composition are also considered, so property changes between the wild-type amino acid and the substituted amino acid residue are

incorporated. In most cases, the backbone atomic composition does not change, except when mutation patterns involve Glycine (Gly), which lacks C_β. This is also duly recorded. For side-chains, we count changes in each of the chemical elements, and obtain an overall atomic compositional change for a given mutation pattern.

2.4 Sequence-Based Features

We employ a standard procedure for multiple sequence alignment (MSA). We first construct the sequence knowledge database from reference proteomes of 1,553 eukaryotic species, excluding that of human. This database contains about 24 millions protein sequences. For each queried human protein, we use BLASTp and Clustal-W2 to obtain its homologs, and select those with identity greater than 30% to the human protein sequence [46, 45]. From the assembled homologous sequences, we construct the MSA using Clustal Omega [47]. We then calculate wild-type frequency, and mutated-type frequency at the mutation site, as well as site-specific entropy, from the aligned MSA using Clustal Omega [47]. Substitution scores are taken from the BLOSUM62 matrix [48].

2.5 Prediction Model and Performance Comparisons

Overall, the features we use for predictions are structural properties, evolutionary signals, and biophysical properties and changes upon substitutions. All are numerical values, except geometric location, which is categorical. We then train a random forest predictor implemented in R [49]. The number of ensemble trees is set to 500, with each tree fit to a balanced training set such that each tree receives an equal number of positive and negative cases. We use 5-fold cross validation on a stratified test dataset for both UnifyPDBFull and UnifyPDBAcceptable datasets [50]. Prediction results by other meth-

ods are retrieved from their open-access web servers or released datasets as of September 18 2021 [10, 11, 12, 14, 13, 15, 28]. Some methods (SPRI, LIST, PMUT, Polyphen-2 and Rhapsody) provide a probability value explicitly, others (EVmutation, FATHMM and PROVEAN) provide raw scores, which are scaled to the range between 0 and 1 using min max normalization [51]. We compare different methods by computing Receiver Operating Characteristic (ROC) curve and Precision-Recall (PR) curve. When no predictions can be made by a specific method, we employ an imputation method to assign a random value from uniform distribution in the range of 0 to 1 to compensate the missing predictions [52].

3 RESULTS

3.1 Benchmark Dataset with Reduced Bias

For reliable prediction of pathological effects and deleterious mutations, a dataset that accurately reflects the natural landscape of mutations is critical. Several benchmark datasets have been constructed in previous studies. They are based on the premises that missense mutations known to result in Mendelian disorders can be regarded as deleterious mutations, whereas mutations with unknown effect can be treated heuristically as neutral mutations. These datasets include HumVar, HumDiv, ExoVar, VariBench, and PredictSNP, and are in wide used [13, 12, 54, 55, 56]. However, these datasets are highly biased, as deleterious mutations and neutral mutations are often from different proteins: there is only a small proportion of proteins possess both deleterious and neutral variants.

There are inherent biases and uncertainties in these datasets, as genes and proteins have varying degree of deleteriousness. In addition, lack of function

knowledge may also introduce biases. First, different proteins have different level of deleteriousness. As previous studies have shown, mutations at sites in functional domains are more likely to have pathological outcomes [57, 58]. In addition, proteins with larger functional surfaces are likely to contain more sites sensitive to substitutions. Hence, proteins with multiple functional domains [59, 60] and larger functional surfaces are less tolerant to mutations, whereas proteins with single function and smaller functional surfaces are more tolerant to mutations. Second, many proteins were not fully investigated, thus variants lacking annotation of pathological effects may be simply a reflection of lack of knowledge, rather than positive information about these mutations being neutral. As a result, a significant drawback of predictions trained using existing datasets is that inconsistent behavior in predictions may result when applied to different types of data. For example, the FATHMM method performs well using the dataset VariBench [11], but exhibits high false positive rate when using the HumDiv dataset [17], indicating that there may be issues of lacks of robustness and transferability.

To overcome such inherent bias, we construct two data sets based on the HumDiv, HumVar and PredictSNP datasets as benchmarks [13, 56]. We select the subset of proteins with annotations of both deleterious and neutral mutations. This is to ensure that the same level of knowledge of both types of mutations exist in the selected proteins. We then select proteins with high-quality structures deposited in the Protein Data Bank (PDB). The smaller dataset, called UnifyPDBFull, contains only proteins with full-coverage structures. The larger dataset, called UnifyPDBAcceptable, contains proteins with both full coverage and non-overlapping partially covered structures. Each protein in these datasets has both deleterious and neutral mutations annotated. Overall, the UnifyPDBFull dataset contains 4,231 deleterious variants and

2,791 neutral variants, which are derived from 252 proteins with 252 polypeptide chains. The UnifyPDBAcceptable dataset contains 5,999 deleterious variants and 3,485 neutral variants, which are derived from 377 proteins with 444 polypeptide chains, indicating the UnifyPDBFull set as a proper subset.

With the more stringent requirement of annotations of knowledge of both deleterious and neutral mutations for each protein, our datasets are smaller compared to other datasets. However, they have much less bias, providing more accountable representations of the landscape of deleterious and neutral mutations in proteins.

3.2 Structural, Biophysical, and Evolutionary Properties for Predicting Mutation Effects

Input features to the machine learning classifier are critical for quantitative prediction of mutation effect, as they should provide adequate information on the structural, biophysical, as well as evolutionary properties of the wild-type residues and changes upon mutations for accurate predictions.

3.2.1 Protein Surface Pocket and Geometric Location of Mutated Residues.

Surface pockets on proteins provide the local microenvironment for ligand binding and biochemical reactions [30, 61]. The buried core may contribute to folding stability [62]. We therefore classify each residue into the categories of surface pockets residues, interior buried residues, and residues on other surface region [37, 34]. Furthermore, we record the solvent accessible surface area of each residue. Pocket construction, surface and buried core residues, as well as solvent accessible surface areas are computed using the alpha shape method

and are accessible from the Computer Atlas of Surface Topography of Proteins (CASTp) server [43, 26].

3.2.2 Contact Profiles for Atomic Interactions in Short Range.

An amino acid residue may play its biochemical roles in conjunction with its spatial neighbors. Together they may form a favorable microenvironment for biochemical reactions to occur. Patterns of such microenvironments are reflected in how side-chains of neighboring residues are arranged and what types of atomic interactions are involved, including both surface and interior residues. In previous studies, profiles of residue distances have been used to report the local environment, which are constructed by recording the frequencies of each amino acid type appearing within certain Euclidean distances between the centriods of the mutation site residue and the neighboring residue [63, 27]. However, important patterns in atomic interactions and side-chain packing arrangement are not fully captured in these distance-based profiles.

Here we construct contact and interaction profiles of mutation site residue incorporating both atomic and residue interaction information. For atomic interactions, we construct the interaction network derived from the weighted alpha shape, which corresponds to the Voronoi diagram with non-empty nearest neighbor atomic contacts. As only those with physical contacts are selected by alpha shape [64], this allows us to focus on exact atomic contacts without the distraction of noise inherent in contacts defined by Euclidean distances.

Specifically, we first computed the weighted Delanuay triangulation of the protein structure. We then obtain its alpha shape by generating a filtration of the Delaunay simplicial complex [64, 33]. We further select alpha edges connecting different residues and obtain inter-residue atomic connection network, where nodes are atoms and alpha edges connect nearest-neighboring atoms whose vol-

ume overlaps. We then organize these contact interactions into profiles of short-range, intermediate-range, and long-range interaction, according to the distance to the mutation site residue measured in the number of alpha edges.

The short-range contact profile is used to capture immediate atomic interaction contributed by interacting residues. For simplicity, we consider only four atom types of carbon (C), nitrogen (N), oxygen (O), and sulfide (S). Atoms provided by the mutation site residue and the neighboring residue are distinguished. For example, carbon from the mutation site and nitrogen from the neighboring residue are recorded as *CN*, and carbon from the neighboring residue and nitrogen from the mutation site are recorded as *NC*, so donor and acceptor information is encoded. We dispense with detailed chemical information of orbital hybridisation of chemical element to avoid overfitting [65, 66]. Altogether, we have 16 types of element pairs for atomic interaction. Furthermore, we record only number counts of atomic interactions and ignore their distance or volume overlap measures as proteins often experience conformational fluctuations [67].

3.2.3 Residue Interaction Profile for Intermediate and Long Range.

There are intermediate and long range interactions in proteins where residues in distance can influence the biochemical environment of the mutation site residue via electrostatic interactions and indirect steric effects [70, 68, 69]. We construct interaction profiles of intermediate- and long-range at the residue level to account for such effects. Interacting residues are again identified by the atomic connection network provided by the alpha shape.

Specifically, we start with residues in direct atom contacts with the mutation site residue, as recorded in the short-range contact profile. We call this the *first layer* contact residues. We then use a Breadth-First Search (BFS) algorithm to identify residues participating in atomic interactions with these *first-layer*

residues [42]. The newly identified residues have *intermediate-range* interactions with the mutation site residue. We continue this process, and identify partner residues that have atomic interactions with residues in the intermediate-range. Those that are newly identified are termed to have *long-range* interactions with the mutation site residue. We then construct the intermediate- and long-range interaction profiles, each a 20-element feature vector recording the number of residues interacting with the mutation site residue of each specific amino acid types.

3.2.4 Salt Bridge Interactions.

It is well-known that non-covalent interactions such as salt bridge plays important roles in protein stability and functions. Salt bridge formed between ion pairs can stabilize protein conformation and provide protons for catalytic reactions [71]. We count the total number of salt bridges that an ionizable residue participates in. A salt bridge forms between the side-chain oxygen atoms in an acidic residue and the side-chain nitrogen atoms in a basic residue if their Euclidean distance is ≤ 3.2 Å.

3.2.5 Changes in Biophysical Properties in Variants.

Substitutions at each mutation site can potentially replace the wild-type residue with any of the 19 other residue types. Different substitutions occurring at the same site may have dramatically different effects. It is therefore important to investigate changes in biophysical properties upon a particular substitution, so differential effects occurring at the same site that are specific to substitution types can be assessed.

To capture changes in biophysical properties, we record differences in occurrence of each atomic chemical type in the side-chain upon mutation. If Gly is involved in the substitution pattern, we record instead the change in C_β in the backbone. This provides information that complements atomic interactions in short-range contact profiles. We further calculate changes in charges upon mutations among ionizable residues and non-ionizable residues. Altogether, these measures allow us to have finer distinction of mutation effects of different substitution patterns at a given mutation site.

3.2.6 Evolutionary Signals.

We use BLOSUM62 substitution matrix to captures general patterns of substitutions during evolution [48]. To incorporate protein-specific evolutionary information, we construct multiple sequence alignments (MSA) following existing methods [27, 10, 12, 13, 15, 14, 16, 28]. Here MSA is constructed using the full protein sequence instead of domains as the structural coverage of proteins is high. Specifically, we use the query human protein to identify homologs with full-length sequence identity $\geq 30\%$ in the reference proteomes of other eukaryotic species. From the assembled homologous sequences, we construct the MSA using CLUSTAL OMEGA [47]. As results, 94.4% (356 out of the 377 UNIFYPDBACCEPTABLE proteins) protein-specific MSAs analyzed in this study have adequate alignment depth of ≥ 100, ensuring that reliable evolutionary signals can be detected. Detailed alignment depth information can be found in Supplemental Information. We then calculate the entropy value for the mutation site using the aligned MSA [47]. In addition, frequencies of occurrence of the wild-type residue and the mutated residue type in the MSA for each mutation site are also recorded.

Our approach is different from that of other methods, in which evolutionary

analysis is carried out at the domain level, which requires significant alignment depth and prior knowledge on functional domain. As a result, difficulties are often encountered when sequence information of annotated functional domains is sparse [10]. Many mutated residues located at those domains without annotation are not covered by alignments, therefore no pathogenicity predictions can be made [10]. Furthermore, we eliminate *ad-hoc* re-weighting or other post-processings methods, unlike the practice of the LIST and POLYPHEN-2 methods [13, 12].

3.3 Evaluating Pathological Effects of Missense Mutations

Information extracted from protein structures, biophysical properties, and evolution analysis are then integrated to generate a probability score that measures the pathological effects of a particular substitution. We use a random forest classifier for this task [49], it is robust and suffers less from overfitting as each tree in the forest is regularized by training on a random subset of the data and a limited subset of features at each branch point [38]. The prediction score π_{del} is calculated from the ratio of the number of trees supporting the hypothesis that the variant is pathological against the total number of trees [38]. It is a method to be proven effective in evaluating mutation effects, as shown in the RHAPSODY and the PMUT studies [16, 28, 14].

3.4 SPRI Performs Well in Predicting Pathogenic Effects on Mendelian Mutations

Table 1: Performance of Predictions on the UnifyPDBFull 5-Fold Test Datasets

	Our method SPRI	EVmutation	FATHMM	LIST	PMUT	Polyphen-2	PROVEAN	Rhapsody
Completeness	**1.0**	0.740	0.987	0.987	0.990	0.987	0.987	0.930
Recall	**0.919**	0.908	0.868	0.755	0.732	0.875	0.842	0.826
Specificity	0.826	0.767	0.385	0.744	0.703	0.813	0.816	**0.926**
Precision	0.889	0.861	0.677	0.814	0.788	0.875	0.872	**0.947**
Accuracy	**0.882**	0.854	0.674	0.75	0.721	0.85	0.832	0.864
F-1 Score	**0.904**	0.884	0.761	0.783	0.759	0.875	0.857	0.882
MCC Score	**0.752**	0.688	0.293	0.492	0.429	0.688	0.654	0.732

Bold indicates the best metric

Table 2: Performance on the UnifyPDBAcceptable 5-Fold Test Datasets

	Our Method SPRI	EVmutation	FATHMM	LIST	PMUT	Polyphen-2	PROVEAN	Rhapsody
Completeness	**1.0**	0.711	0.990	0.990	0.987	0.990	0.990	0.941
Recall	**0.913**	0.902	0.846	0.753	0.747	0.881	0.824	0.822
Specificity	0.828	0.766	0.405	0.766	0.732	0.809	0.817	**0.925**
Precision	0.902	0.878	0.707	0.846	0.829	0.888	0.884	**0.952**
Accuracy	**0.882**	0.854	0.683	0.758	0.742	0.855	0.821	0.859
F-1 Score	**0.907**	0.890	0.771	0.797	0.786	0.884	0.853	0.882
MCC Score	**0.744**	0.676	0.281	0.505	0.466	0.689	0.628	0.721

Bold indicates the best metric

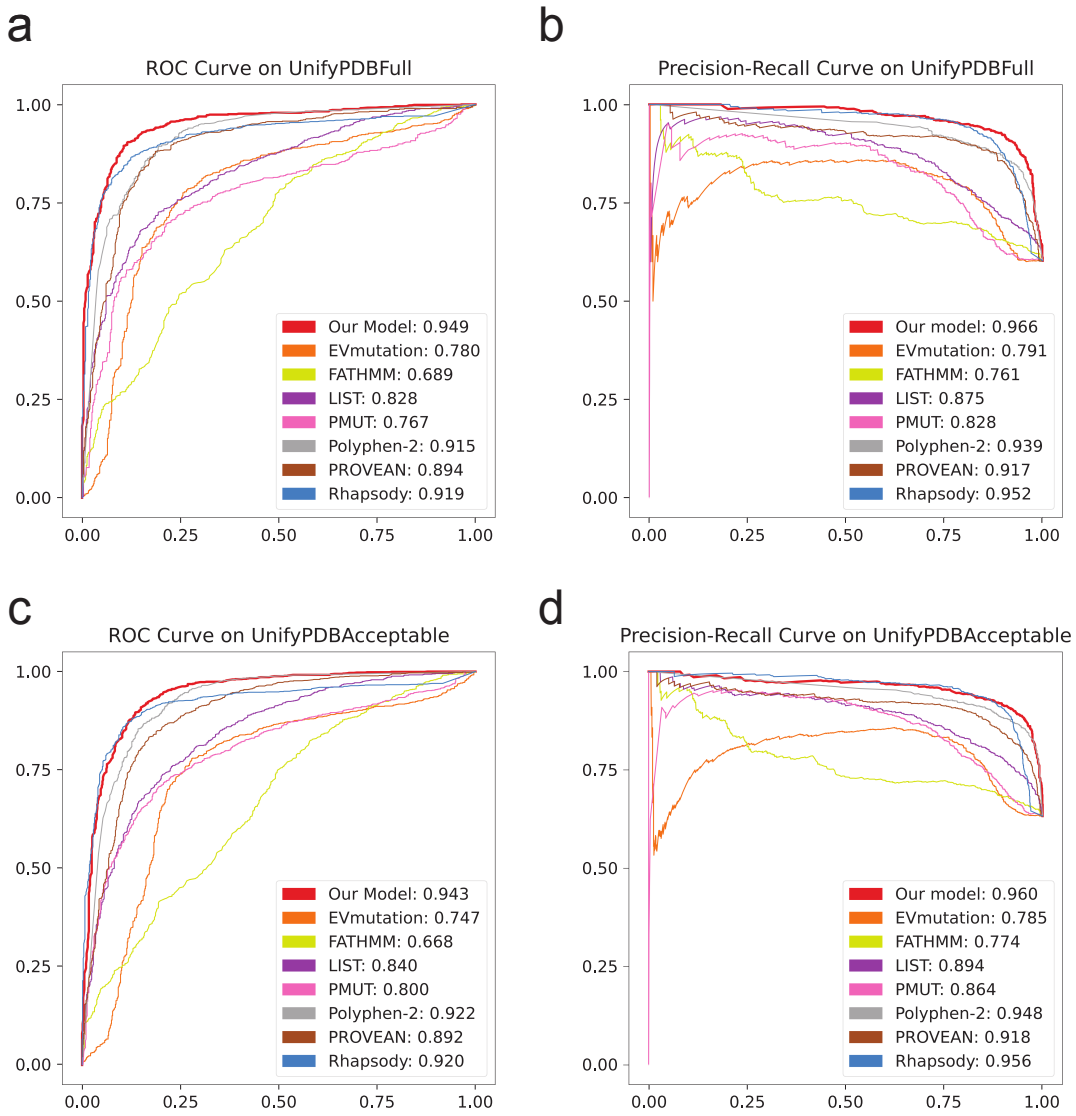

Figure 2: Receiver Operating Characteristic (ROC) Curve and Precision-Recall (PR) Curve of testing results using SPRI and other methods on the UnifyPDBFull and UnifyPDBAcceptable test datasets. (a) The ROC curve using the UnifyPDBFull test dataset. SPRI has the highest AU-ROC of 0.949. (b) The PR curve using the UnifyPDBFull test dataset. SPRI has the highest AU-PRC of 0.966. (c) The ROC curve using the UnifyPDBAcceptable test dataset. SPRI has the highest AU-ROC of 0.943. (d) The PR curve using the UnifyPDBAcceptable test dataset. SPRI has the highest AU-PRC of 0.960.

To evaluate how well our method works, we use stratified k-fold cross-validations on both the UnifyPDBFull and the UnifyPDBAcceptable datasets [50]. We focus on the UnifyPDBFull dataset, since it contains proteins whose structures covers the full sequences, ensuring accurate representations of their structural properties. We compare our results with those from the methods of EVmutation, FATHMM, LIST, PMUT, Polyphen-2, PROVEAN and Rhaposdy [10, 11, 12, 14, 13, 15, 28]. We use default thresholds provided by FATHMM, PMUT, Polyphen-2, PROVEAN and Rhaposdy to distinguish deleterious mutations from neutral mutations [11, 14, 13, 15]. We use the optimal thresholds for EVmutation, LIST, and our method at the highest Matthews correlation coefficient (MCC) values for evaluation.

We report completeness, recall, specificity, precision, accuracy, F-1 score, and MCC for comprehensiveness. Among these metrics, MCC is an informative and balanced metric to symmetrically evaluates both groups of positives and negatives with adjustment of ratio of subgroups, hence provides a good metric on summarizing prediction results [72]. Overall, our method SPRI has excellent performance in distinguishing deleterious missense mutations from neutral mutations. The detailed metric values are listed in Table 1.

The 5-fold average performance of our method on UnifyPDBFull stratified test data has an F-1 score and MCC of 0.904 and of 0.752, respectively, which are the highest among all 8 methods compared. Our method also has the highest accuracy (0.882) and highest recall (0.919), indicating it is highly sensitive in identifying deleterious mutations. In addition, our method has an excellent specificity of 0.826 in identifying neutral mutations correctly, which are better than all other methods except Rhapsody (0.926). Our methods outperforms Rhapsody in several important regards: Rhapsody has a lower recall of 0.826, and it can make predictions on less variants in UnifyPDBFull (0.930),

whereas our SPRI method has a higher recall of 0.919 and makes predictions on all variants.

We also compute the Receiver Operating Characteristic (ROC) and the Precision-Recall (PR) curves to assess the robustness of the performance of prediction at different thresholds, which are showing in Figure 2. Both ROC curve and PR curve are used to examine performance for binary classification models, with PR curve providing a more informative assessment of imbalanced dataset. On the stratified UnifyPDBFull test dataset, our method has an AUC-ROC of 0.946 and an AUC-PRC of 0.966, which are the best among all methods.

These results demonstrate that our method is robust and performs well at different thresholds with different settings. Results using the other cross-validation data (UnifyPDBAcceptable) are similar to that of UnifyPDB-Full, where our method has the best performance when measured by the metrics of F1-score, MCC value, accuracy, and AUCs of ROC and PR curves.

We also identify features that are most important for accurate prediction (see Supplemental Information). Features from all of the three categories (structural properties, evolutionary signals from sequence analysis, and biophysical properties) play important roles in distinguishing deleterious mutations from neutral mutations. These results suggest that our features capture essential biological properties that when mutationally disturbed can contribute to pathogenesis.

Overall, our method has the best performance by most measurements among the 8 methods compared, and is robust in making accurate predictions at different thresholds, despite the fact that our model was derived on a training data set of much smaller size.

3.5 Transferability in Identifying Cancer Driver Mutations

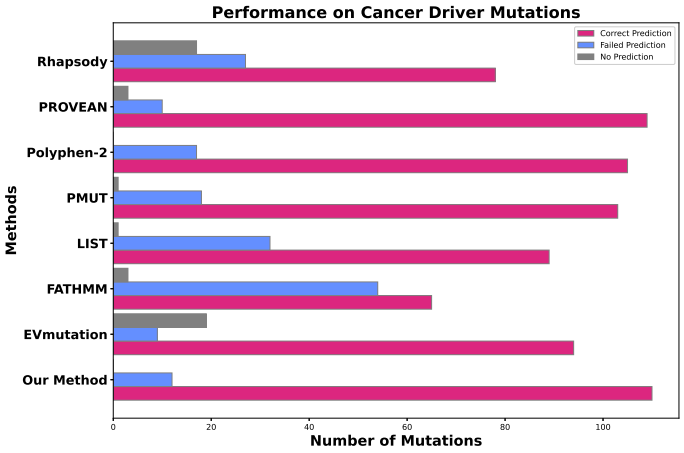

Figure 3: Predictions of cancer driver mutations among the set of CANCER CENSUS MUTATIONS (CMC) tier-1 mutations. For each method, the numbers of correct prediction (pink), failed prediction (light blue), and missing predictions (gray) are listed. Overall, our SPRI method has the highest number of 110 correct predictions out of 122 driver mutations.

While SPRI has been developed and evaluated using germline mutations of Mendelian disorders, we hypothesize that our method predicting mutation effects trained on Mendelian disorders captures the essential characteristics of pathological mutations and is transferable to evaluate effects of cancer somatic mutations.

It is well known that somatic mutations provide one of the most important means to trigger abnormal cell growth that may lead to tumorigenesis [73, 74, 75]. Identifying somatic mutations that drive cancer development is important, as it facilitates development of therapeutics targeting these mutated proteins. For example, sotorasib has been recently approved to treat lung cancer patients with KRAS G12C mutation [76]. However, it is challenging to identify cancer driver mutations, as there are millions of missense mutations accumulated in cancer patients, the majority of which exhibit low recurrence [6]. Approaches based on frequency counting are therefore not effective.

To test our hypothesis, we examine whether our method can be used to identify cancer driver mutations. For this task, we take the recently available annotations of tier 1 CANCER CENSUS MUTATION from the COSMIC V92 as the ground truth on cancer driver mutations [6]. These mutations are designated by COSMIC as there is strong evidence to support their roles in tumorigenesis from experimental validation, clinical evidence, and *in silico* analysis [6]. This dataset contains 201 missense mutations containing both wild-type residue and substituted amino acid. Among these, 122 mutations can be mapped to 50 mutated residues from 26 proteins with known structures. For these cancer driver mutations, we build a separate predictor following the same procedure but with the 26 cancer relevant proteins excluded from the training process. We use the same probability threshold for classification.

The evaluation results show that our method has excellent performance in

identifying cancer driver mutations. Our method can correctly classify 110 mutations out of the 122 structure-mapped mutations as deleterious. We also tested the methods of PROVEAN, Polyphen-2 and PMUT, which have comparable performance, identifying 109, 105, and 103 deleterious mutations, respectively. EVmutation identifies 94 deleterious mutations, with 9 incorrect predictions and 19 missing predictions. Rhapsody identifies 78 out of the 122 deleterious mutations, at a significantly reduced level of effectiveness than Mendelian disorders. FATHMM identifies 65 deleterious mutations, with 54 incorrect predictions and 3 missing predictions.

Overall, our method has excellent performance in identifying confirmed cancer driver mutations. These results demonstrate that our method captures essential properties of disease-causing mutations. It can provide out-of-the-box functionality to evaluate pathological effects on more complex mutations in cancer, without additional training, parameter tuning, or additional prior knowledge.

3.6 A Case Study on Human Glutathione Synthetase

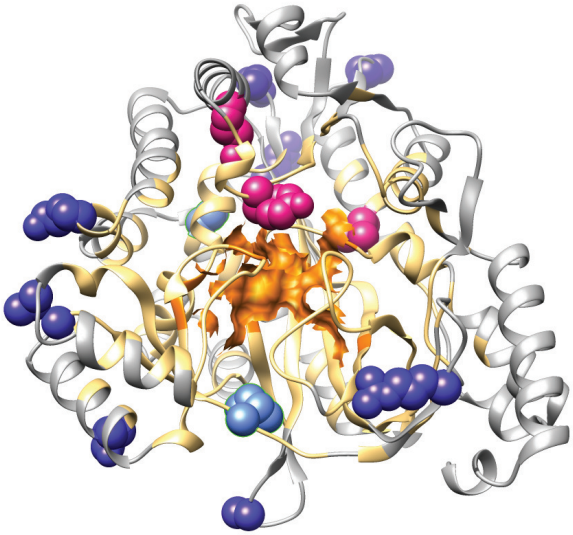

Figure 4: The catalytic region, its neighboring environment, and outside region of the human glutathione synthetase. The surface of catalytic region is colored in orange, the neighboring environment in light orange, and the outside region in grey. The correctly predicted deleterious mutations are in pink. The correctly predicted neutral mutations are in dark blue. and the false predictions of deleterious mutations are in light blue. All 4 deleterious mutation site residues (L188P, Y270C, Y270H, R283C) are located within the neighboring environment of the catalytic region, and are correctly predicted by SPRI. Eight correctly predicted neutral predictions are located in the outside region, and the two false deleterious predictions are located at the boundary of the catalytic neighboring environment.

To illustrate how our method can be useful for gaining mechanistic insight, here we discuss detailed analysis of human glutathione synthetase. Glutathione synthetase catalyses the conversion of gamma-L-glutamyl-L-cysteine and glycine to phosphate and glutathione in the presence of ATP. The human glutathione synthetase (UniProt: P48637, PDB: `2HGS` chain:A) has 524 residues, 9 of which form its catalytic region, with Glu144, Asn146 and Glu368 forming the magnesium binding site, and Arg125, Ser151, Lys305, Lys364, Gly369 and Arg450 play the role of electrostatic stabilizer [77]. There are 4 deleterious variants occurring at 3 mutation sites and 10 neutral variants occurring at 10 different mutation sites in the UnifyPDBFull dataset. Among these, the variant R283C leads to glutathione synthetase deficiency symptom [78]. Our method provides reliable predictions on these variants. Specifically, all 4 deleterious variants (L188P, Y270C, Y270H, R283C) are correctly predicted to be deleterious, 8 out of 10 neutral variants (K95E, A134T, P202T, H290C, V343M, R418Q, Q435H, E353K) are correctly predicted to be neutral, and only 2 neutral variants (S80N, I401T) are incorrectly predicted to be deleterious.

The spatial relationship among deleterious and neutral variants and the interaction profiles of functional residues provides useful insight. We take the 9 residues which form the catalytic region as the center, and compute the 3-layer interaction profiles using alpha-shape. This profile defines the neighboring environment of the catalytic region. All other residues removed from the catalytic region and its neighboring environment are recorded as outside residues. Overall, we found that all deleterious variants are located within the neighboring environment of the catalytic region, 8 neutral variants with correct predictions are outside this region. The two neutral variants with incorrect predictions are at the boundary of the neighboring environment of the catalytic region, suggesting that the interaction profile for sites at the boundaries of the catalytic

region may require further refinement.

We have collected a total of 19 well-studied enzymes from the UnifyPDB-Full dataset that have accurate information on catalytic residues according to the manually annotated M-CSA enzyme database [79]. We follow the same procedure to define the catalytic region, its neighboring environment, and the outside region. The null hypothesis is that deleterious and neutral mutations exhibit no difference in preference for locations neighboring the enzyme catalytic regions. The alternative hypothesis is that deleterious variants are more preferably located at the catalytic or its neighboring region, while neutral variants are more likely to occur at the outside region. The p-value of 4.2×10^{-14} from Fisher's exact test strongly rejects the null hypothesis. These results indicate that the novel interaction profiles can reliably capture information on important residues contributing to protein functions. Furthermore, it is important to note that unlike the detailed analysis presented in this section, our prediction method does not require any prior knowledge on enzyme catalytic residues.

3.7 Predictions for Mutation Saturation

As SPRI can predict deleterious mutations contributing to both Mendelian disorders and complex diseases such as cancer, it can be extended to assess general effects of substitutions of arbitrary sites from the wild type to any of the other 19 amino acid types, if the structure of the protein is known. Fig 5 depicts an example, where we show the heatmap of pathological effects of all different substitutions at all positions along the full protein sequence of the copper transport protein. The color intensity of each square in the heatmap represents the level of pathological effect of a substitution. Our method is scalable, and such heatmaps can be computed proteome wide.

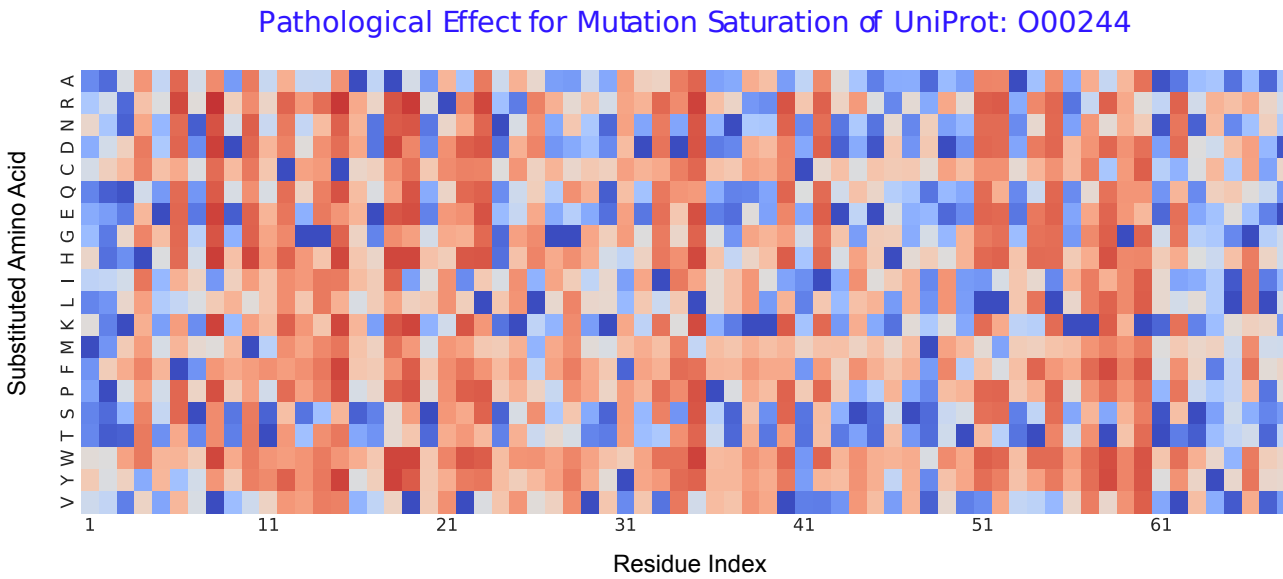

Figure 5: The heatmap of predicted pathological effects of mutation saturation of copper transport protein ATOX1 (UNIPROT: O00244). The horizontal axis are the residue index along the sequence. The vertical axis represents the substituted amino acid types. Each square is color coded by the effects of missense mutation, where the color intensity encodes the likelihood of pathological effect. A red substitution represents a mutations predicted to be most likely a pathological mutation, whereas a blue substitution represents a mutation predicted to be most likely a neutral mutation.

4 DISCUSSION and CONCLUSION

As large-scale variant data of exon regions become widely available, computational methods identifying missense mutations that adversely affect biological functions are essential for interpretation and for understanding these variant data. An important source of valuable information for this task is the 3D structures of proteins. With the rapid expansion of experimental determination [18, 21, 20, 19] and computational prediction of protein structures [80, 81], this additional source of information will be available for most proteins of interests. Recent studies showed that important advances beyond sequence-based analysis can be made when structural properties [27] and structure-derived dy-

namics [16, 28] are considered. Nevertheless, a number of widely-used methods are based on sequence analysis alone [10, 11, 12, 14, 13, 15] and do not take advantage of structural information.

However, how to effectively decipher structural information determining biochemical functions remains challenging. In this study, we report the novel method SPRI that incorporates properties of explicitly computed surface pockets and other geometric and topological properties of the protein structures. This approach is motivated by the fact that biochemical functions require specific spatial arrangement of residues and atoms to present a binding surface with required biochemical properties [30]. Through geometric computation of both atomic and residue-level of interactions of short, intermediate, and long ranges, as well as incorporation of biophysical properties upon missense mutations, our SPRI method exhibits strong performance in identifying deleterious mutations and in quantifying pathogenicity in Mendelian disease, which compares favorably to current state-of-the-art methods.

Our method also sheds light on an important question, namely, whether pathological missense mutations of different varieties are organized by the same principles, regardless of germline or somatic origin, and whether it manifests as Mendelian diseases or complex diseases. Previous studies mostly have built disease-specific pathogenicity models. For example, FATHMM has different predictors for inherited disease, cancer, and other specific pathologies [11]. PMUT allows user-input customized training datasets to be used so different predictors for different diseases can be constructed [14]. There are drawbacks with these approaches. FATHMM has very different answers in evaluating pathogenic mutations, depending on whether a predictor for inherited diseases or a predictor for cancer is used. This is likely due to the different level of available annotated information and the adjustment of thresholds [11]. While

Rhapsody works well in predicting Mendelian disease related mutations, its performance in predicting cancer driver mutations lags behind significantly [28]. Therefore it is unclear if overall biochemical principles behind different disease mutations exist. If so, they are not well-captured by current methods. Our results presented here suggest different types of pathological missense mutations are all governed by the same biochemical principles. In addition, the relevant biophysical properties required are largely encoded in the protein structure and sequence. Furthermore, they are effectively extracted by SPRI. We show accurate predictions can be made on cancer driver mutations, even though SPRI was trained using Mendelian disease mutations and there were no parameter adjustment or re-training using different data.

Our results suggest that the relevant biophysical properties can be effectively extracted from protein structures, robustly for analysis using different PDB structures of the same protein (see Supplemental Information). This is illustrated by the analysis of glutathione synthetase, where it is shown that deleterious variants are more likely to occur in the neighborhood of the catalytic regions, while neutral variants likely occur outside the functional regions. These neighborhoods can be precisely defined using surface analysis and atomic-residue interactions profiles. It is important to note that such extraction can be made without a *prior* knowledge of functional annotation of specific residues, indicating characteristics of the functional binding surfaces are encoded in the protein structures and sequences, and are effectively extracted by our method.

There are limitations in our method. Knowledge of the protein structures is required for assessing effects of the missense mutations. In fact, our results show that predictions on proteins with knowledge of the full structures are better than those on proteins with partial structures: The MCC score on UnifyPDBAcceptable dataset containing incomplete structures is slightly inferior than the

MCC score on UnifyPDBFull dataset containing full structures. We anticipate that as additional structural information become available from techniques such as cryo-EM or AlphaFold2 [19, 80], this issue will be resolved as more proteins will have their sequence fully covered by structures. The broad availability of most protein structures will also allow us to construct a global atlas of saturation mutation maps, similar to the example of the copper transport protein illustrated in this study. Such an atlas can provide global views of protein fitness landscape and mutation effects of the whole protein universe.

Another limitation of our method is currently it is based on analysis of protein structure of monomeric units. Structures of protein complex contains valuable information on protein-protein interactions. As structures of protein-complexes continue to increase [82] and technical issues [83] being resolved, we anticipate that incorporating knowledge of protein-protein interaction interfaces will further improve models pathogenicity of missense mutations.

5 ACKNOWLEDGEMENTS

This work is supported by NIH grants R35 GM127084.

5.1 Conflict of interest statement.

None declared.

References

[1] Yang, Y., Muzny, D., Xia, F., Niu, Z., Person, R., Ding, Y., Ward, P., Braxton, A., Wang, M., Buhay, C. & Others Molecular findings among patients referred for clinical whole-exome sequencing. *Jama*. **312**, 1870-1879 (2014)

[2] Ng, P. & Kirkness, E. Whole genome sequencing. *Genetic Variation*. pp. 215-226 (2010)

[3] Belkadi, A., Bolze, A., Itan, Y., Cobat, A., Vincent, Q., Antipenko, A., Shang, L., Boisson, B., Casanova, J. & Abel, L. Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. *Proceedings Of The National Academy Of Sciences*. **112**, 5473-5478 (2015)

[4] Lek, M., Karczewski, K., Minikel, E., Samocha, K., Banks, E., Fennell, T., O'Donnell-Luria, A., Ware, J., Hill, A., Cummings, B. & Others Analysis of protein-coding genetic variation in 60,706 humans. *Nature*. **536**, 285-291 (2016)

[5] Consortium, 1. & Others A map of human genome variation from population scale sequencing. *Nature*. **467**, 1061 (2010)

[6] Tate, J., Bamford, S., Jubb, H., Sondka, Z., Beare, D., Bindal, N., Boutselakis, H., Cole, C., Creatore, C., Dawson, E. & Others COSMIC: the catalogue of somatic mutations in cancer. *Nucleic Acids Research*. **47**, D941-D947 (2019)

[7] Labelle, Y., Phaneuf, D., Leclerc, B. & M. Tanguay, R. Characterization of the human fumarylacetoacetate hydrolase gene and identification of a missense mutation abolishing enzymatic activity. *Human Molecular Genetics*. **2**, 941-946 (1993)

[8] Kawaguchi, T., Kato, S., Otsuka, K., Watanabe, G., Kumabe, T., Tominaga, T., Yoshimoto, T. & Ishioka, C. The relationship among p53 oligomer formation, structure and transcriptional activity using a comprehensive missense mutation library. *Oncogene*. **24**, 6976-6981 (2005)

[9] Vaser, R., Adusumalli, S., Leng, S., Sikic, M. & Ng, P. SIFT missense predictions for genomes. *Nature Protocols*. **11**, 1-9 (2016)

[10] Hopf, T., Ingraham, J., Poelwijk, F., Schärfe, C., Springer, M., Sander, C. & Marks, D. Mutation effects predicted from sequence co-variation. *Nature Biotechnology*. **35**, 128-135 (2017)

[11] Shihab, H., Gough, J., Cooper, D., Stenson, P., Barker, G., Edwards, K., Day, I. & Gaunt, T. Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Human Mutation*. **34**, 57-65 (2013)

[12] Malhis, N., Jones, S. & Gsponer, J. Improved measures for evolutionary conservation that exploit taxonomy distances. *Nature Communications*. **10**, 1-8 (2019)

[13] Adzhubei, I., Schmidt, S., Peshkin, L., Ramensky, V., Gerasimova, A., Bork, P., Kondrashov, A. & Sunyaev, S. A method and server for predicting damaging missense mutations. *Nature Methods*. **7**, 248-249 (2010)

[14] López-Ferrando, V., Gazzo, A., De La Cruz, X., Orozco, M. & Gelpı, J. PMut: a web-based tool for the annotation of pathological variants on proteins, 2017 update. *Nucleic Acids Research*. **45**, W222-W228 (2017)

[15] Choi, Y. & Chan, A. PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. *Bioinformatics*. **31**, 2745-2747 (2015)

[16] Ponzoni, L. & Bahar, I. Structural dynamics is a determinant of the functional significance of missense variants. *Proceedings Of The National Academy Of Sciences*. **115**, 4164-4169 (2018)

[17] Wang, B., Tian, W., Lei, X., Perez-Rathke, A., Tseng, Y. & Liang, J. Structure-based Method for Predicting Deleterious Missense SNPs. *2019 IEEE EMBS International Conference On Biomedical & Health Informatics (BHI)*. pp. 1-4 (2019)

[18] Rose, Y., Duarte, J., Lowe, R., Segura, J., Bi, C., Bhikadiya, C., Chen, L., Rose, A., Bittrich, S., Burley, S. & Others RCSB Protein Data Bank: Architectural advances towards integrated searching and efficient access to macromolecular structure data from the PDB archive. *Journal Of Molecular Biology*. **433**, 166704 (2021)

[19] Yip, K., Fischer, N., Paknia, E., Chari, A. & Stark, H. Atomic-resolution protein structure determination by cryo-EM. *Nature*. **587**, 157-161 (2020)

[20] Wüthrich, K. Protein structure determination in solution by NMR spectroscopy.. *Journal Of Biological Chemistry*. **265**, 22059-22062 (1990)

[21] Ilari, A. & Savino, C. Protein structure determination by x-ray crystallography. *Bioinformatics*. pp. 63-87 (2008)

[22] Sedova, M., Iyer, M., Li, Z., Jaroszewski, L., Post, K., Hrabe, T., Porta-Pardo, E. & Godzik, A. Cancer3D 2.0: interactive analysis of 3D patterns of cancer mutations in cancer subsets. *Nucleic Acids Research*. **47**, D895-D899 (2019)

[23] Meyer, M., Lapcevic, R., Romero, A., Yoon, M., Das, J., Beltrán, J., Mort, M., Stenson, P., Cooper, D., Paccanaro, A. & Others mutation3D: cancer

gene prediction through atomic clustering of coding variants in the structural proteome. *Human Mutation.* **37**, 447-456 (2016)

[24] Chen, S., He, X., Li, R., Duan, X. & Niu, B. HotSpot3D web server: an integrated resource for mutation analysis in protein 3D structures. *Bioinformatics.* **36**, 3944-3946 (2020)

[25] Ryslik, G., Cheng, Y., Cheung, K., Bjornson, R., Zelterman, D., Modis, Y. & Zhao, H. A spatial simulation approach to account for protein structure when identifying non-random somatic mutations. *BMC Bioinformatics.* **15**, 1-13 (2014)

[26] Stitziel, N., Binkowski, T., Tseng, Y., Kasif, S. & Liang, J. topoSNP: a topographic database of non-synonymous single nucleotide polymorphisms with and without known disease association. *Nucleic Acids Research.* **32**, D520-D522 (2004)

[27] Quan, L., Wu, H., Lyu, Q. & Zhang, Y. DAMpred: Recognizing Disease-Associated nsSNPs through Bayes-Guided Neural-Network Model Built on Low-Resolution Structure Prediction of Proteins and Protein–Protein Interactions. *Journal Of Molecular Biology.* **431**, 2449-2459 (2019)

[28] Ponzoni, L., Peñaherrera, D., Oltvai, Z. & Bahar, I. Rhapsody: Predicting the pathogenicity of human missense variants. *Bioinformatics.* **36**, 3084-3092 (2020)

[29] Bakan, A., Meireles, L. & Bahar, I. ProDy: protein dynamics inferred from theory and experiments. *Bioinformatics.* **27**, 1575-1577 (2011)

[30] English, A., Groom, C. & Hubbard, R. Experimental and computational mapping of the binding surface of a crystalline protein. *Protein Engineering.* **14**, 47-59 (2001)

[31] Sigala, P., Fafarman, A., Bogard, P., Boxer, S. & Herschlag, D. Do ligand binding and solvent exclusion alter the electrostatic character within the oxyanion hole of an enzymatic active site?. *Journal Of The American Chemical Society.* **129**, 12104-12105 (2007)

[32] Liang, J., Woodward, C. & Edelsbrunner, H. Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. *Protein Science.* **7**, 1884-1897 (1998)

[33] Edelsbrunner, H. & Mücke, E. Three-dimensional alpha shapes. *ACM Transactions On Graphics (TOG).* **13**, 43-72 (1994)

[34] Tseng, Y., Dundas, J. & Liang, J. Predicting protein function and binding profile via matching of local evolutionary and geometric surface patterns. *Journal Of Molecular Biology.* **387**, 451-464 (2009)

[35] Tseng, Y. & Liang, J. Estimation of amino acid residue substitution rates at local spatial regions and application in protein function inference: a Bayesian Monte Carlo approach. *Molecular Biology And Evolution.* **23**, 421-436 (2006)

[36] Dundas, J., Adamian, L. & Liang, J. Structural signatures of enzyme binding pockets from order-independent surface alignment: a study of metalloendopeptidase and NAD binding proteins. *Journal Of Molecular Biology.* **406**, 713-729 (2011)

[37] Liang, J., Edelsbrunner, H., Fu, P., Sudhakar, P. & Subramaniam, S. Analytical shape computation of macromolecules: I. Molecular area and volume through alpha shape. *Proteins: Structure, Function, And Bioinformatics.* **33**, 1-17 (1998)

[38] Breiman, L. Random forests. *Machine Learning.* **45**, 5-32 (2001)

[39] Consortium, U. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Research.* **47**, D506-D515 (2019)

[40] Wang, B., Lei, X., Tian, W., Perez-Rathke, A., Tseng, Y. & Liang, J. SeqMapPDB: A Standalone Pipeline to Identify Representative Structures of Protein Sequences and Mapping Residue Indices in Real-Time at Proteome Scale. *ArXiv Preprint ArXiv:2202.11551.* (2022)

[41] Li, X., Hu, C. & Liang, J. Simplicial edge representation of protein structures and alpha contact potential with confidence measure. *Proteins: Structure, Function, And Bioinformatics.* **53**, 792-805 (2003)

[42] Bundy, A. & Wallen, L. Breadth-first search. *Catalogue Of Artificial Intelligence Tools.* pp. 13-13 (1984)

[43] Tian, W., Chen, C., Lei, X., Zhao, J. & Liang, J. CASTp 3.0: computed atlas of surface topography of proteins. *Nucleic Acids Research.* **46**, W363-W367 (2018)

[44] Phillips, J., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., Chipot, C., Skeel, R., Kale, L. & Schulten, K. Scalable molecular dynamics with NAMD. *Journal Of Computational Chemistry.* **26**, 1781-1802 (2005)

[45] Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. & Madden, T. BLAST+: architecture and applications. *BMC Bioinformatics.* **10**, 1-9 (2009)

[46] Larkin, M., Blackshields, G., Brown, N., Chenna, R., McGettigan, P., McWilliam, H., Valentin, F., Wallace, I., Wilm, A., Lopez, R. & Others Clustal W and Clustal X version 2.0. *Bioinformatics.* **23**, 2947-2948 (2007)

[47] Sievers, F. & Higgins, D. Clustal omega. *Current Protocols In Bioinformatics.* **48**, 3-13 (2014)

[48] Henikoff, S. & Henikoff, J. Amino acid substitution matrices from protein blocks. *Proceedings Of The National Academy Of Sciences.* **89**, 10915-10919 (1992)

[49] Liaw, A., Wiener, M. & Others Classification and regression by randomForest. *R News.* **2**, 18-22 (2002)

[50] Zeng, X. & Martinez, T. Distribution-balanced stratified cross-validation for accuracy estimation. *Journal Of Experimental & Theoretical Artificial Intelligence.* **12**, 1-12 (2000)

[51] Jain, A., Nandakumar, K. & Ross, A. Score normalization in multimodal biometric systems. *Pattern Recognition.* **38**, 2270-2285 (2005)

[52] Lin, W. & Tsai, C. Missing value imputation: a review and analysis of the literature (2006–2017). *Artificial Intelligence Review.* **53**, 1487-1509 (2020)

[53] Landrum, M., Lee, J., Benson, M., Brown, G., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Jang, W. & Others ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Research.* **46**, D1062-D1067 (2018)

[54] Karczewski, K., Weisburd, B., Thomas, B., Solomonson, M., Ruderfer, D., Kavanagh, D., Hamamsy, T., Lek, M., Samocha, K., Cummings, B. & Others The ExAC browser: displaying reference data information from over 60 000 exomes. *Nucleic Acids Research.* **45**, D840-D845 (2017)

[55] Nair, P. & Vihinen, M. V ari B ench: A benchmark database for variations. *Human Mutation.* **34**, 42-49 (2013)

[56] Bendl, J., Stourac, J., Salanda, O., Pavelka, A., Wieben, E., Zendulka, J., Brezovsky, J. & Damborsky, J. PredictSNP: robust and accurate consensus

classifier for prediction of disease-related mutations. *PLoS Computational Biology.* **10**, e1003440 (2014)

[57] Guo, H., Choe, J. & Loeb, L. Protein tolerance to random amino acid change. *Proceedings Of The National Academy Of Sciences.* **101**, 9205-9210 (2004)

[58] Kaminker, J., Zhang, Y., Watanabe, C. & Zhang, Z. CanPredict: a computational tool for predicting cancer-associated missense mutations. *Nucleic Acids Research.* **35**, W595-W598 (2007)

[59] Gussow, A., Petrovski, S., Wang, Q., Allen, A. & Goldstein, D. The intolerance to functional genetic variation of protein domains predicts the localization of pathogenic mutations within genes. *Genome Biology.* **17**, 1-11 (2016)

[60] Mani, M., Chen, C., Amblee, V., Liu, H., Mathur, T., Zwicke, G., Zabad, S., Patel, B., Thakkar, J. & Jeffery, C. MoonProt: a database for proteins that are known to moonlight. *Nucleic Acids Research.* **43**, D277-D282 (2015)

[61] Billas, I., Iwema, T., Garnier, J., Mitschler, A., Rochel, N. & Moras, D. Structural adaptability in the ligand-binding pocket of the ecdysone hormone receptor. *Nature.* **426**, 91-96 (2003)

[62] Gromiha, M., Oobatake, M., Kono, H., Uedaira, H. & Sarai, A. Relationship between amino acid properties and protein stability: buried mutations. *Journal Of Protein Chemistry.* **18**, 565-578 (1999)

[63] Pires, D., Ascher, D. & Blundell, T. mCSM: predicting the effects of mutations in proteins using graph-based signatures. *Bioinformatics.* **30**, 335-342 (2014)

[64] Edelsbrunner, H. Shape reconstruction with Delaunay complex. *Latin American Symposium On Theoretical Informatics*. pp. 119-132 (1998)

[65] Kirtman, B., Chipman, D. & Palke, W. Orbital hybridization. *Journal Of The American Chemical Society*. **99**, 1305-1307 (1977)

[66] Zhao, B., Tam, Y. & Zheng, J. An autoencoder with bilingual sparse features for improved statistical machine translation. *2014 IEEE International Conference On Acoustics, Speech And Signal Processing (ICASSP)*. pp. 7103-7107 (2014)

[67] Damjanovich, S., Somogyi, B. & Welch, G. Protein fluctuation and enzyme activity. *Journal Of Theoretical Biology*. **105**, 25-33 (1983)

[68] Shoulders, M., Hodges, J. & Raines, R. Reciprocity of steric and stereoelectronic effects in the collagen triple helix. *Journal Of The American Chemical Society*. **128**, 8112-8113 (2006)

[69] Baldridge, A., Samanta, S., Jayaraj, N., Ramamurthy, V. & Tolbert, L. Steric and electronic effects in capsule-confined green fluorescent protein chromophores. *Journal Of The American Chemical Society*. **133**, 712-715 (2011)

[70] Nakamura, H. Roles of electrostatic interaction in proteins. *Quarterly Reviews Of Biophysics*. **29**, 1-90 (1996)

[71] Kumar, S. & Nussinov, R. Salt bridge stability in monomeric proteins. *Journal Of Molecular Biology*. **293**, 1241-1255 (1999)

[72] Chicco, D. & Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*. **21**, 1-13 (2020)

[73] Greenman, C., Stephens, P., Smith, R., Dalgliesh, G., Hunter, C., Bignell, G., Davies, H., Teague, J., Butler, A., Stevens, C. & Others Patterns of somatic mutation in human cancer genomes. *Nature*. **446**, 153-158 (2007)

[74] Ramroop, J., Gerber, M. & Toland, A. Germline variants impact somatic events during tumorigenesis. *Trends In Genetics*. **35**, 515-526 (2019)

[75] Wang, M., Zhao, J., Zhang, L., Wei, F., Lian, Y., Wu, Y., Gong, Z., Zhang, S., Zhou, J., Cao, K. & Others Role of tumor microenvironment in tumorigenesis. *Journal Of Cancer*. **8**, 761 (2017)

[76] Hong, D., Fakih, M., Strickler, J., Desai, J., Durm, G., Shapiro, G., Falchook, G., Price, T., Sacher, A., Denlinger, C. & Others KRASG12C inhibition with sotorasib in advanced solid tumors. *New England Journal Of Medicine*. **383**, 1207-1217 (2020)

[77] Dinescu, A., Cundari, T., Bhansali, V., Luo, J. & Anderson, M. Function of conserved residues of human glutathione synthetase: implications for the ATP-grasp enzymes. *Journal Of Biological Chemistry*. **279**, 22412-22421 (2004)

[78] Nj, R., Carlsson, K., Bhansali, V., Luo, J., Nilsson, L., Ladenstein, R., Anderson, M., Larsson, A. & Norgren, S. Human hereditary glutathione synthetase deficiency: kinetic properties of mutant enzymes. *Biochemical Journal*. **381**, 489-494 (2004)

[79] Ribeiro, A., Holliday, G., Furnham, N., Tyzack, J., Ferris, K. & Thornton, J. Mechanism and Catalytic Site Atlas (M-CSA): a database of enzyme reaction mechanisms and active sites. *Nucleic Acids Research*. **46**, D618-D623 (2018)

[80] Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žıdek, A., Potapenko, A. & Others Highly accurate protein structure prediction with AlphaFold. *Nature*. **596**, 583-589 (2021)

[81] Du, Z., Su, H., Wang, W., Ye, L., Wei, H., Peng, Z., Anishchenko, I., Baker, D. & Yang, J. The trRosetta server for fast and accurate protein structure prediction. *Nature Protocols*. **16**, 5634-5651 (2021)

[82] Kundrotas, P., Kotthoff, I., Choi, S., Copeland, M. & Vakser, I. Dockground tool for development and benchmarking of protein docking procedures. *Protein Structure Prediction*. pp. 289-300 (2020)

[83] Sprinzak, E., Sattath, S. & Margalit, H. How reliable are experimental protein–protein interaction data?. *Journal Of Molecular Biology*. **327**, 919-923 (2003)