Supporting Information

S1 Datasets of non-synonymous SNVs and their structural coverage

In order to evaluate the impact of various types of non-synonymous variants on localized frustration in different biological contexts, we collected and analyzed data from a variety of sources. These sources were chosen in order to obtain both benign and disease-associated SNVs, with the disease-associated SNVs having been further sub-classified to investigate their associated modes of action in greater detail. An overview of this data collection scheme is provided in ***S1A***. ***S1B*** gives summary statistics on all non-synonymous SNVs in these datasets, and ***S1C*** provides the corresponding data on the subset of these SNVs. Further details on the statistics obtained as part of this data collection framework are provided below.

We collected and annotated 6.46 million non-synonymous SNVs using VAT. About 5.1 million of these SNVs were benign mutations that were obtained from the ExAC Project, and an additional ~0.6 million SNVs were taken from phase 3 of the 1000 Genomes Project, which constitutes 79% and 9% of our total set of annotated SNVs, respectively (***S1B***). The remaining SNVs were a set of disease-associated mutations, and these comprised ~76,000 HGMD SNVs and 0.65 million publicly available pan-cancer somatic SNVs. The HGMD and the pan-cancer dataset constitutes 2% and 10% of our total collected non-synonymous SNVs, respectively (Figure ***S1B***).

However, when considering only those annotated SNVs that map to high-resolution protein structures, the contribution of SNVs from different sources changes significantly. Approximately 96,000 SNVs from ExAC were mapped to protein structures in the PDB, which thus constitute 51% of our total set of structurally mapped SNVs (Figure ***S1C***). Similarly, 1000 Genomes SNVs constitute 7% (13,588) of the total structurally mapped SNVs. In contrast, the percentage of the disease-associated SNVs that may be mapped to high-quality crystal structures was 18% (33,261 SNVs) and 24% (44,094 SNVs) for the HGMD and pan-cancer resource, respectively (Figure ***S1C***). The majority of SNVs from the pan-cancer dataset that were mapped to protein structures impacted cancer-associated genes (CAG), constituting 14% (25,409) of the all structurally-mapped SNVs, whereas SNVs impacting non-cancer associated genes

constitute only 8% (15,044; Figure ***S1C***). In contrast, 4,041 SNVs affecting driver genes may be mapped to protein structures; these SNVs constitute 2% of the total structurally mapped non-synonymous SNVs (Figure ***S1C***).

S2 ΔF distributions within the semi-balanced SNV dataset

We also performed comparisons between 1000 Genomes, ExAC and HGMD variants using the semi-balanced dataset (details in the Method section). We find that, overall, the results obtained using this semi-balanced dataset are consistent with the ΔF distributions obtained above (Supp. Fig S5). However, they lack statistical significance, potentially due to the smaller sample sizes of SNVs included in the semi-balanced set.

S3 Threshold to identify potentially deleterious SNVs

As discussed in the Results section of the main text, disease-associated SNVs from HGMD generally induce more negative ΔF values relative to benign SNVs. Given a newly discovered SNV, is there a specific ΔF threshold that may optimally be used to classify SNVs as benign or deleterious? We address this issue empirically by optimizing a simple function *f(x)* defined by two distributions (1):

$$f(x) = h(x) + e(x)$$

Let ΔF_{HGMD} denote the distribution of ΔF scores induced by HGMD SNVs. *h(x)* is defined to be the difference between the fraction of ΔF_{HGMD} scores less than x ($fract[\Delta F_{HGMD} < x]$) and the fraction of ΔF_{HGMD} scores greater than x ($fract[\Delta F_{HGMD} > x]$):

$$h(x) = fract[\Delta F_{HGMD} < x] - fract[\Delta F_{HGMD} > x]$$

ΔF_{ExAC} is similarly defined for the distribution of ΔF values associated with ExAC SNVs:

$$e(x) = fract[\Delta F_{ExAC} > x] - fract[\Delta F_{ExAC} < x]$$

Note that, in building the distribution of ΔF_{HGMD} values, a random sample of HGMD SNVs was chosen in order to match the number of SNVs in the ΔF_{ExAC} distribution. The x that maximizes the function *f(x)* is taken as the ΔF threshold for predicting whether a newly discovered SNV is deleterious or benign. Using this approach, we find that this ideal threshold takes a value of ΔF = -1.221.

SI Figures

Figure S1: Overview of SNV categories and their relative proportions within the data pool analyzed. *A)* Flowchart representing the different categories and origins of the variants analyzed in this study. A given non-synonymous SNV can be classified as benign or disease-associated on the basis of its provenance (i.e., whether it is taken from 1000 Genomes, ExAC, HGMD or Pan-cancer variant datasets). *B)* Relative proportions of SNVs from various datasets *prior* to mapping SNVs to high-resolution PDB structures. *C)* Relative proportions of SNVs from various datasets *after* mapping SNVs to high-resolution PDB structures.

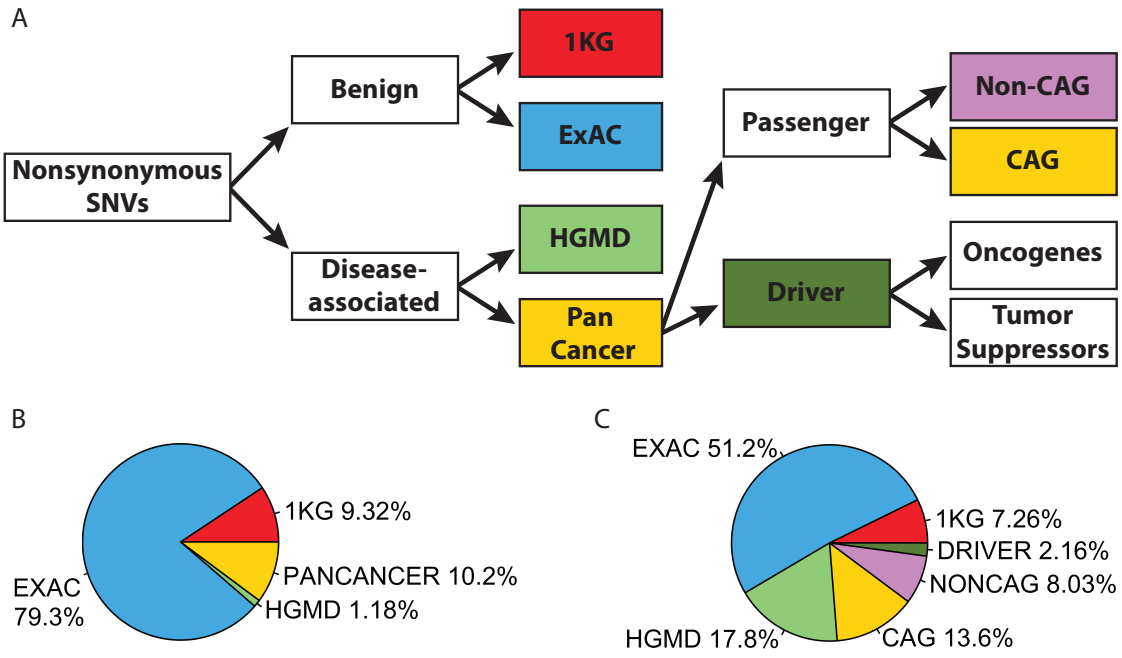


Figure S2: Histogram of the number of 1000 Genomes SNVs against the number of unique proteins. The histogram depicts the distribution of the number of distinct proteins in which non-synonymous 1000 Genomes SNVs may be mapped to high-quality crystal structures within the PDB. A total of 618 distinct proteins are available. Redundancy was removed by ensuring that no pair of proteins within this dataset shares more than 90% sequence identity.

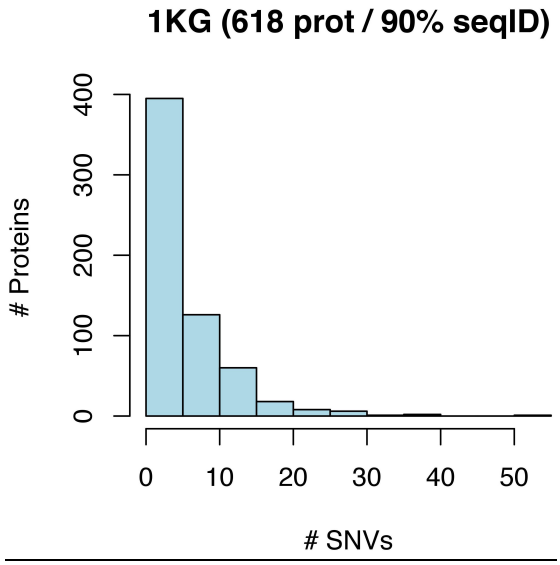


Figure S3: Histogram of the number of ExAC SNVs against the number of unique proteins. The histogram depicts the distribution of the number of distinct proteins in which non-synonymous ExAC SNVs may be mapped to high-quality crystal structures within the PDB. A total of 907 distinct proteins are available. Redundancy was removed by ensuring that no pair of proteins within this dataset shares more than 90% sequence identity.

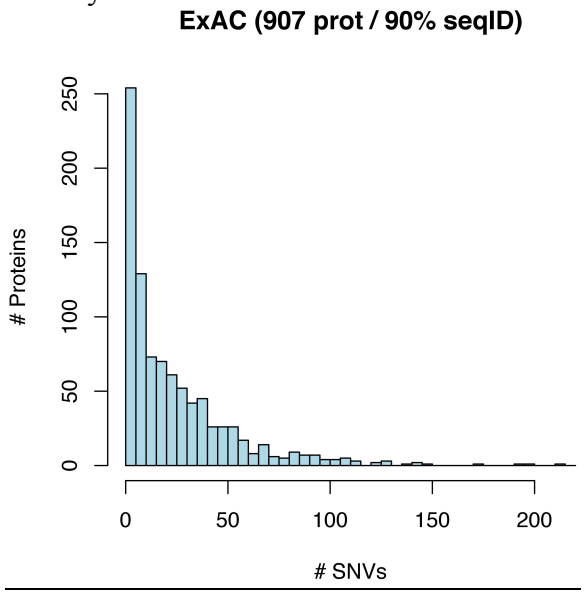


Figure S4: Histogram of the number of HGMD SNVs against the number of unique proteins. The histogram depicts the distribution of the number of distinct proteins in which non-synonymous HGMD SNVs may be mapped to high-quality crystal structures within the PDB. A total of 293 distinct proteins are available. Redundancy was removed by ensuring that no pair of proteins within this dataset shares more than 90% sequence identity.

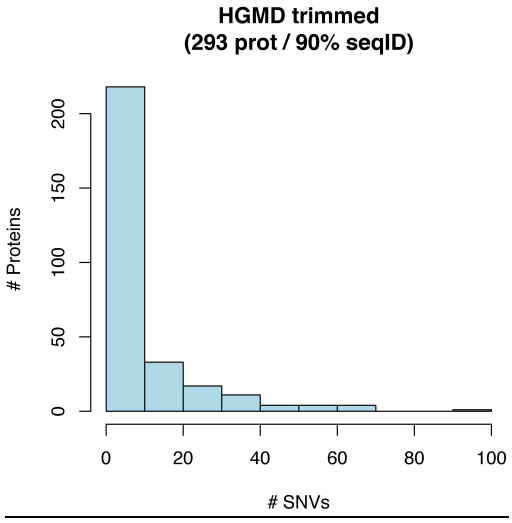


Figure S5: Comparisons between the ΔF distributions within the semi-balanced set of structures. Violin plots showing ΔF distributions associated with SNVs affecting core or surface residues of structures for which at least one SNV is taken from *A)* 1000 Genomes & HGMD, *B)* ExAC & HGMD and *C)* HGMD & ExAC. These trends on the semi-balanced SNV dataset are consistent with observations reported in the main text. However, the smaller sample sizes within the semi-balanced set may result in poorer statistical significance. The white dots, black boxes and vertical lines represent the medians, interquartile ranges, and 95% confidence intervals of the ΔF distributions, respectively.

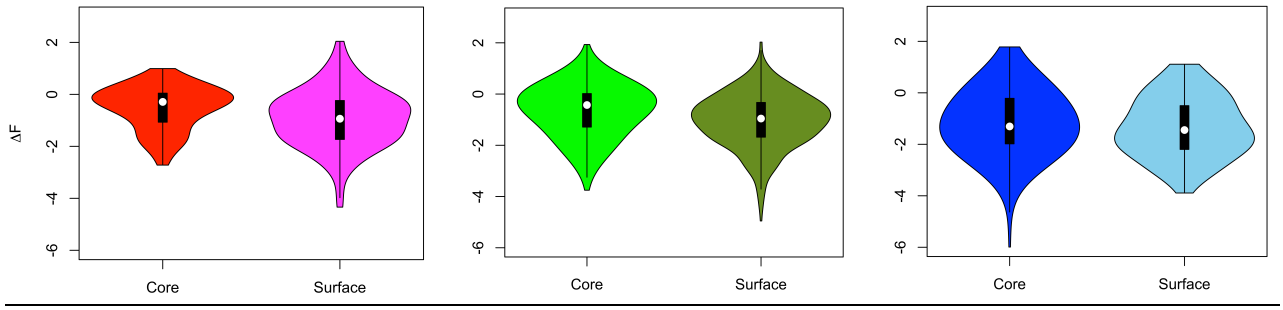


Figure S6: Empirical distribution to predict deleterious SNVs. In order to determine an optimal threshold for discriminating between benign and deleterious SNVs using ΔF, we use a simple function to be optimized with respect to ΔF. Details of the simple formalism used are provided in SI section "S3: Threshold to identify potentially deleterious SNVs". The optimum ΔF value obtained (ΔF = -1.221) is marked with a vertical dotted line. The blue density plot designates the ΔF values associated with benign SNVs from ExAC, and the red density plot designates the ΔF values associated with deleterious SNVs from HGMD. Both plots are normalized to have an integral of unity.

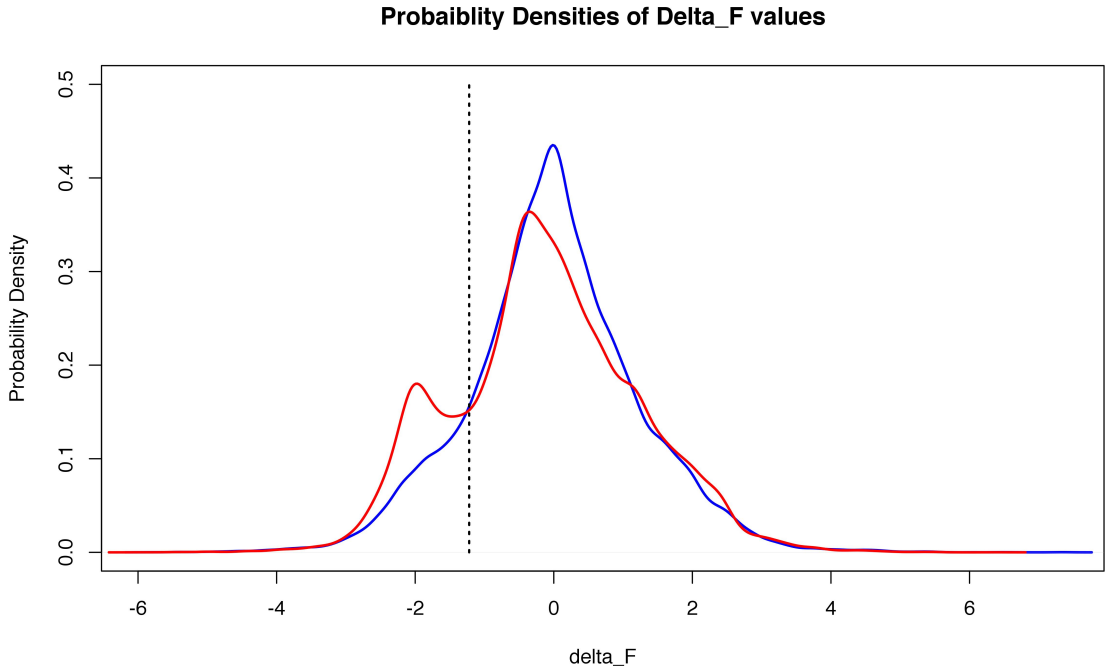
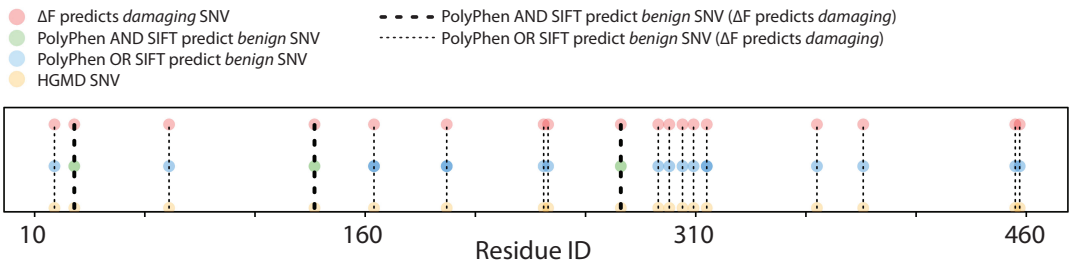


Figure S7: Linearized depiction of HGMD SNVs that constitute ΔF-rescued false negatives. Shown is a linear depiction of the distribution of HGMD SNVs (orange) predicted to be deleterious using a ΔF cutoff of -1.221, along with predictions from PolyPhen and SIFT. Heavy dotted lines demarcate loci in which both PolyPhen and SIFT provide false negatives (i.e., fail to predict deleteriousness), whereas light dotted lines designate SNVs for which either PolyPhen-2 or SIFT provide false negatives. The particular structure shown corresponds to human glucokinase (PDB ID: 1V4S).



Reference

1. Hourai Y, Akutsu T, Akiyama Y (2004) Optimizing substitution matrices by separating score distributions. *Bioinformatics* 20(6):863–73.