Table S1. Training dataset composition

	|Toggle|	|Neutral|	|Rheostat|	SNP-Possible	|Unknown|	|Filtered|	Details
raw	-	-	-	-	822	-	
X-residues					820	2	undetermined amino acid X
SNP-possible	-	-	-	591	229	-	
k-means labeling	66	152	-	372	229	-	k-means clustering exp. scores
manual refinement	60	147	-	372	229	12	removed: T(6), N(6)
ntModel labels	94	181	151	61	113	-	removed: T(74), N(109), R(48)
manual refinement	20	72	104	-	-	231	
funtrpTraining	80	219	104	61	113	245	Final funtrpTraining = 403

Detailed overview of the of training dataset composition for training of both random forest-based models in the *funtrp* pipeline. Show are the total numbers of instances remaining after the cluster labeling, filtering and prediction steps.

Table S2. Set of sequence-based features used to train *funtrp* randomf forest-based. models

id	Feature	Source	Description	Parameters
1	Solvent Accessibility	PROF (*)	predicted solvent accessibility (PACC)	PredictProtein defaults
2	Secondary Structure	PROF (*)	predicted helix (pH), strand (pE) or loop (pL)	PredictProtein defaults
3	Residue Flexibility	PROFbval (*)	predicted residue flexibility (PROFbval)	PredictProtein defaults
4	Protein Disorder	MD (*)	predicted protein disorder (MDraw)	PredictProtein defaults
5	Amino Acid	-	amino acids encoded as a vector of length 20	NA
6	Residue Size	-	basic amino acid property (small or large)	NA
7	Residue Charge	-	basic amino acid property (uncharged / + / -)	NA
8	SNP possible	-	number of possible nsSNPs (all codons)	NA
9	Conservation	ConSurf (*)	predicted conservation	PredictProtein defaults
10	MSA Ratio	-	Total fraction of residue amino acid at MSA column	NA

(*) tools are applied via the PredictProtein pipeline {Yachdav, 2014 #6}. Features were ranked by importance towards fuNTRp position type labels in Swiss-Prot using ReliefF; weights were rounded off {Kononenko, 1996 #30}. If applicable, parameters used in feature computation are specified.

Table S4. *Protein subsets for analysis*

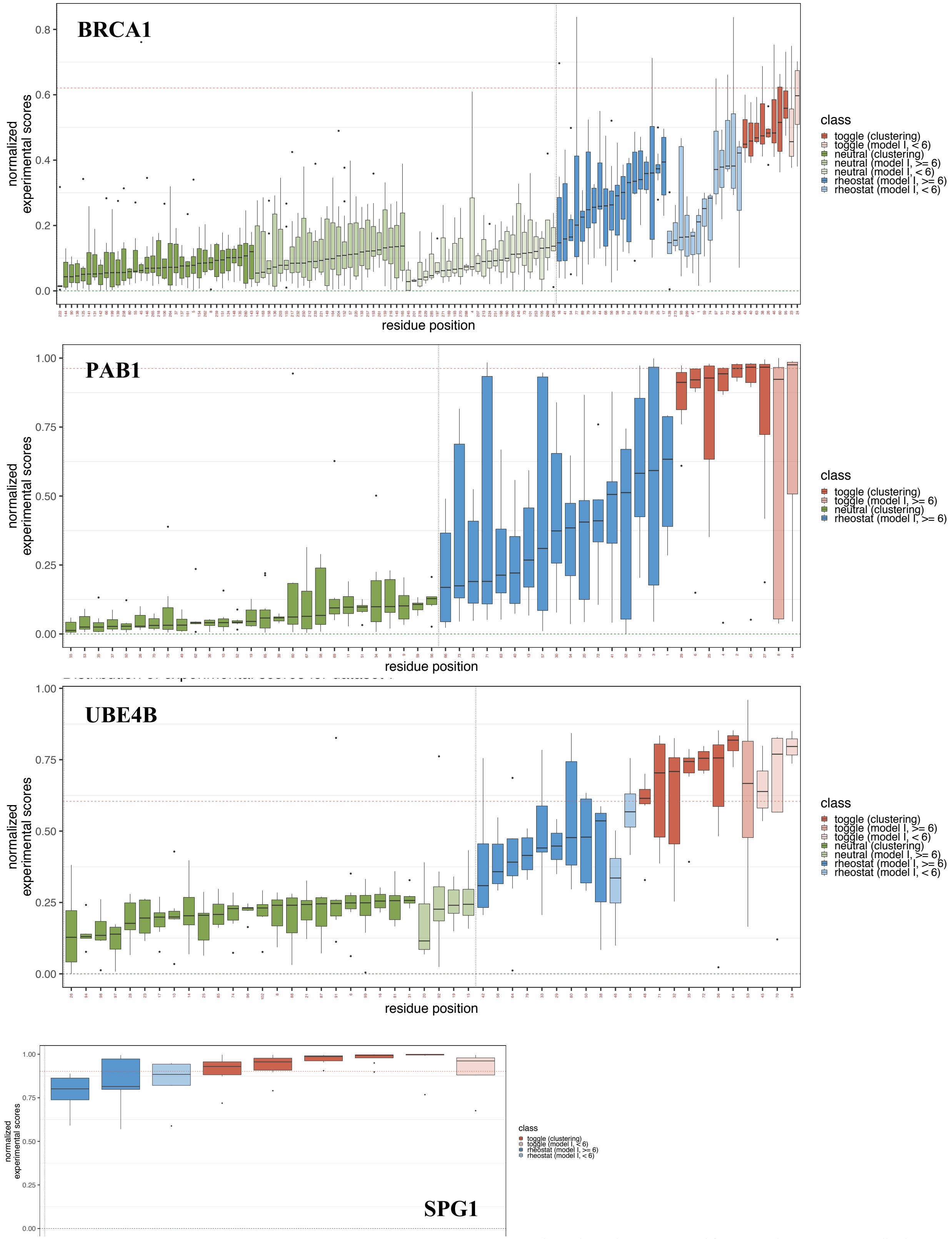

Fig. S3. Distribution of experimental (DMS) variant effect scores for training datasets. Measured experimental scores extracted from DMS datasets were normalized to [0,1]. Residue positions on the x-Axis are grouped by (i) position types, (ii) way of labeling and (iii) within these groupings ordered based on increasing distribution medians. The labeling types are: clustering, predicted with more than six experimental scores available and predicted with less than six experimental scores available.

Identifier	Source	|Proteins| <> |fuNTRp|	|w/ E.C. annotation|	|w/o E.C. annotation|
Swiss-Prot	UniProtKB/SwissProt	20,410 <> 19,501	4.273 <> 4.241	16.137 <> 15.260
TrEMBL	UniProtKB/TrEBML		9.668 <> 9.554	144.277 <> 5.254
EXPV	UniProtKB/SwissProt	1.250 <> 1.239	1.250 <> 1.239	x
PMD	SNPdbe	3.127 <> 3.098	x	x

Extracted datasets used in analysis. EXPV is a subset of experimentally verified enzymes in Swiss-Prot (Mahlich, et al., 2018). Literature based annotations of effect were extracted from the PMD dataset.

Table S5. Confusion matrices of position type predictions for (A) *ntModel* und (B) *funtrpModel*

(A)

Neutral	Toggle	Observed ↓
140	7	Neutral
9	51	Toggle

(B)

Neutral	Toggle	Rheostat	Observed ↓
199	4	16	Neutral
2	64	14	Toggle
19	5	80	Rheostat

Predictions for both models are based on LOO-CV results.

Table S6. Performance of predicting position types for a Random Forest (RF) based classifier model using evolutionary conservation alone

	Sensitivity	Specificity	Pos Pred Value	Neg Pred Value	Precision	Recall	F1	Prevalence	Detection Rate	Detection Prevalence	Balanced Accuracy
N	0.6621509	0.8674369	0.7514676	0.8073990	0.7514676	0.6621509	0.7028646	0.3788741	0.2504617	0.3333009	0.7647939
R	0.4605779	0.7003374	0.2884851	0.8302884	0.2884851	0.4605779	0.3530603	0.2092962	0.0961263	0.3331653	0.5804576
T	0.6568313	0.8926389	0.8098981	0.7873813	0.8098981	0.6568313	0.7247737	0.4118298	0.2701228	0.3335338	0.7747351

Shown are the averaged performances per class over 100 resample runs. For each run, 3000 residue positions from Swiss-Prot were resampled randomly (without replacement), selecting 1000 instances of each position type respectively. The same was repeated for the test set and a total of 300 residue positions. Position type labes were based on *funtrp* predictions.

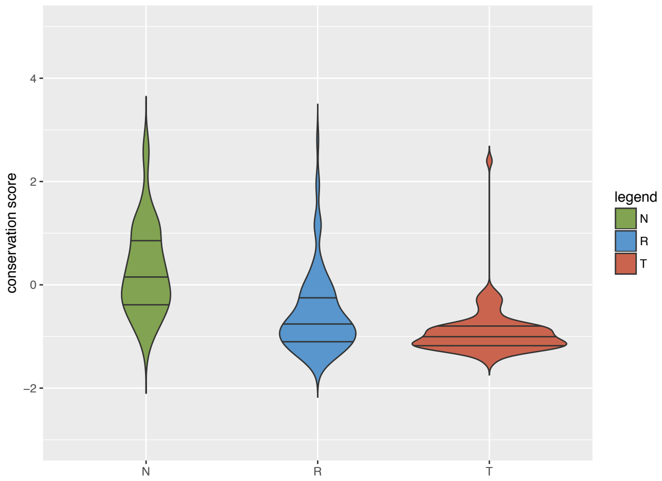

Fig. S7. Distribution of ConSurf conservation scores for fuNTRp training dataset. Density distributions of evolutionary conservation (ConSurf) compared between position types for the *funtro* model training dataset. ConSurf predictions scores are by default normalized such as 0 depicts the average score over the entire protein and standard devia-tion is |1|).

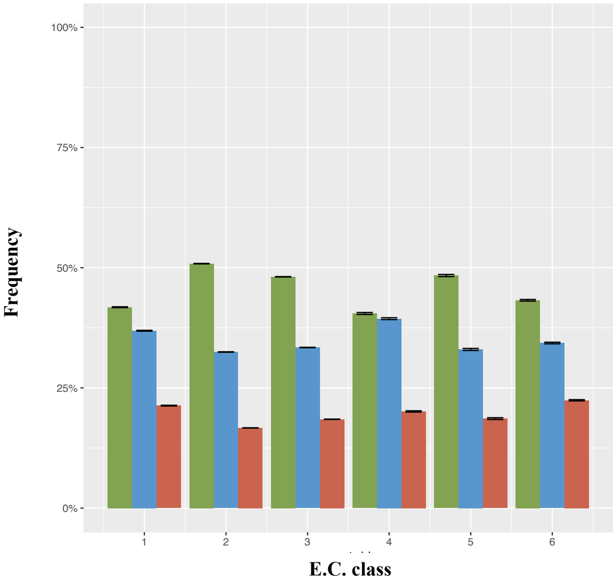

Fig. S8. Distribution of position types for main E.C. classes in the entire Siwss-Prot dataset. *Colors are according to position type (green =Neutral, red =Rheostat, blue =Toggle). Error bars are computed based on 100 iterations of random subsampling (Methods),*

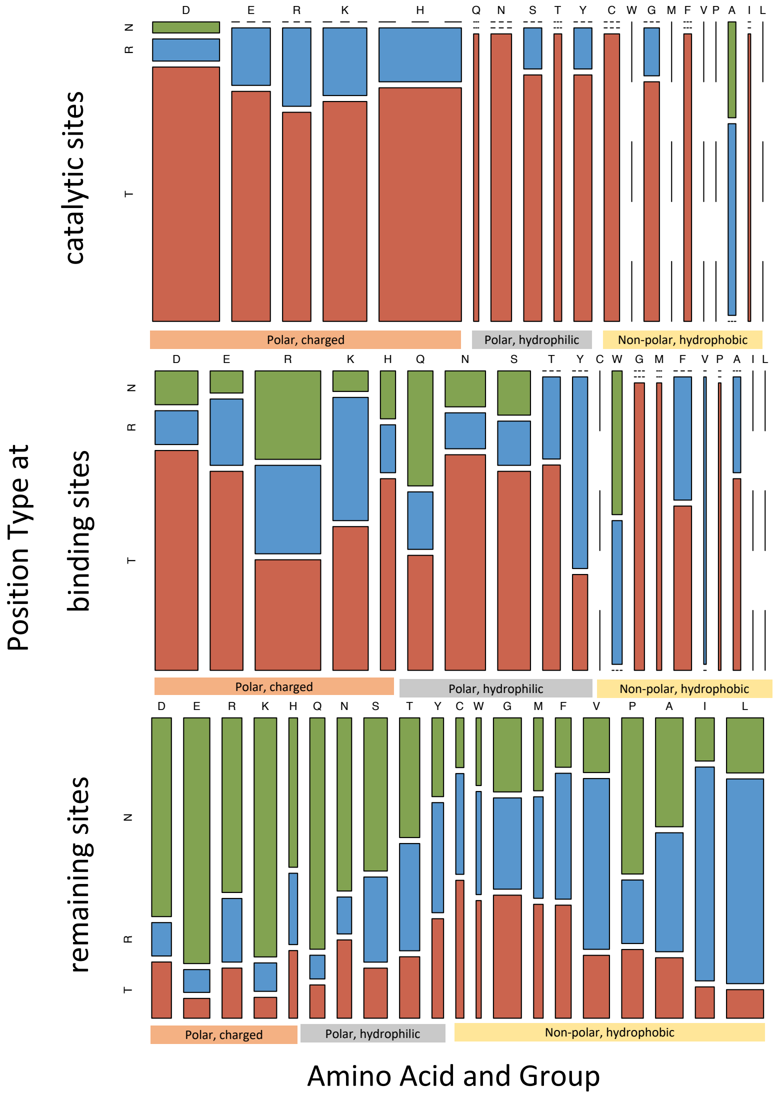

Fig. S9. Fractions of position types per amino acid compared by site characteristic. Comparison of fractions at catalytic sites and binding sites against the remaining residues of the respective Swiss-Prot enzymes. Colors are according to position type (green =Neutral, red =Rheostat, blue =Toggle).

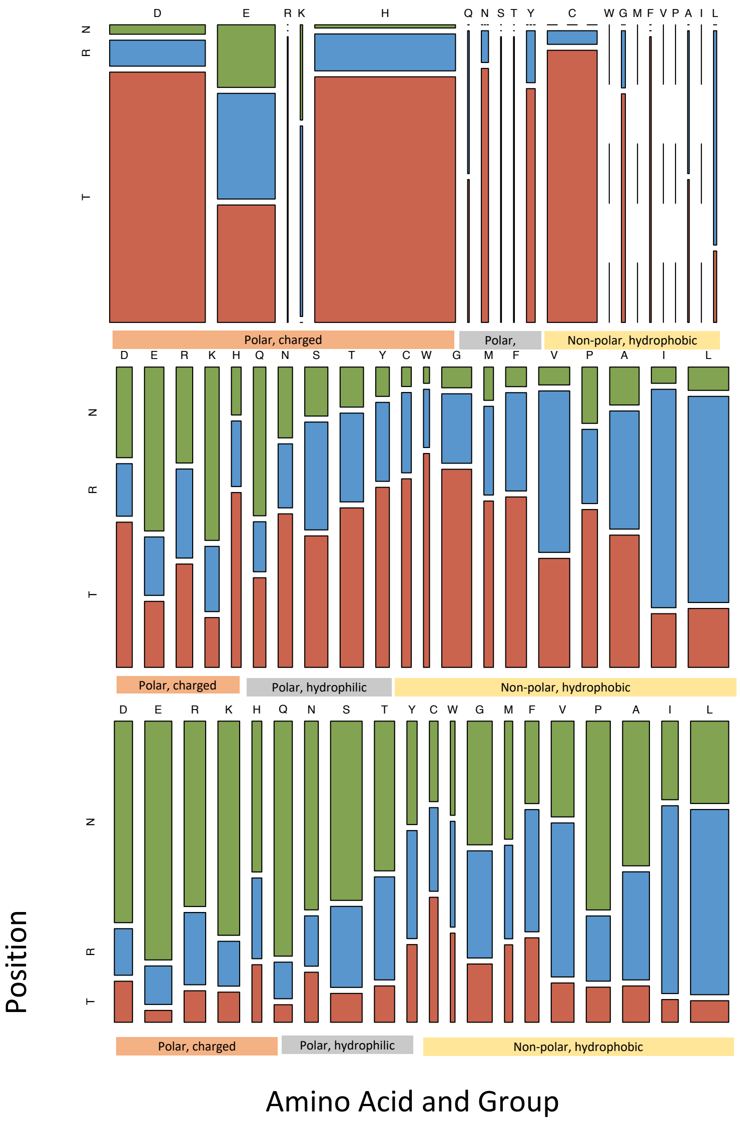

Fig. S10. Fractions of position types per amino acid for metal binding sites and spheres. Comparison of SaHLe spheres and residues annotated as metal binding sites within spheres vs remaining residues of the respective Swiss-Prot enzymes. Colors are according to position type (green =Neutral, red =Rheostat, blue =Toggle).

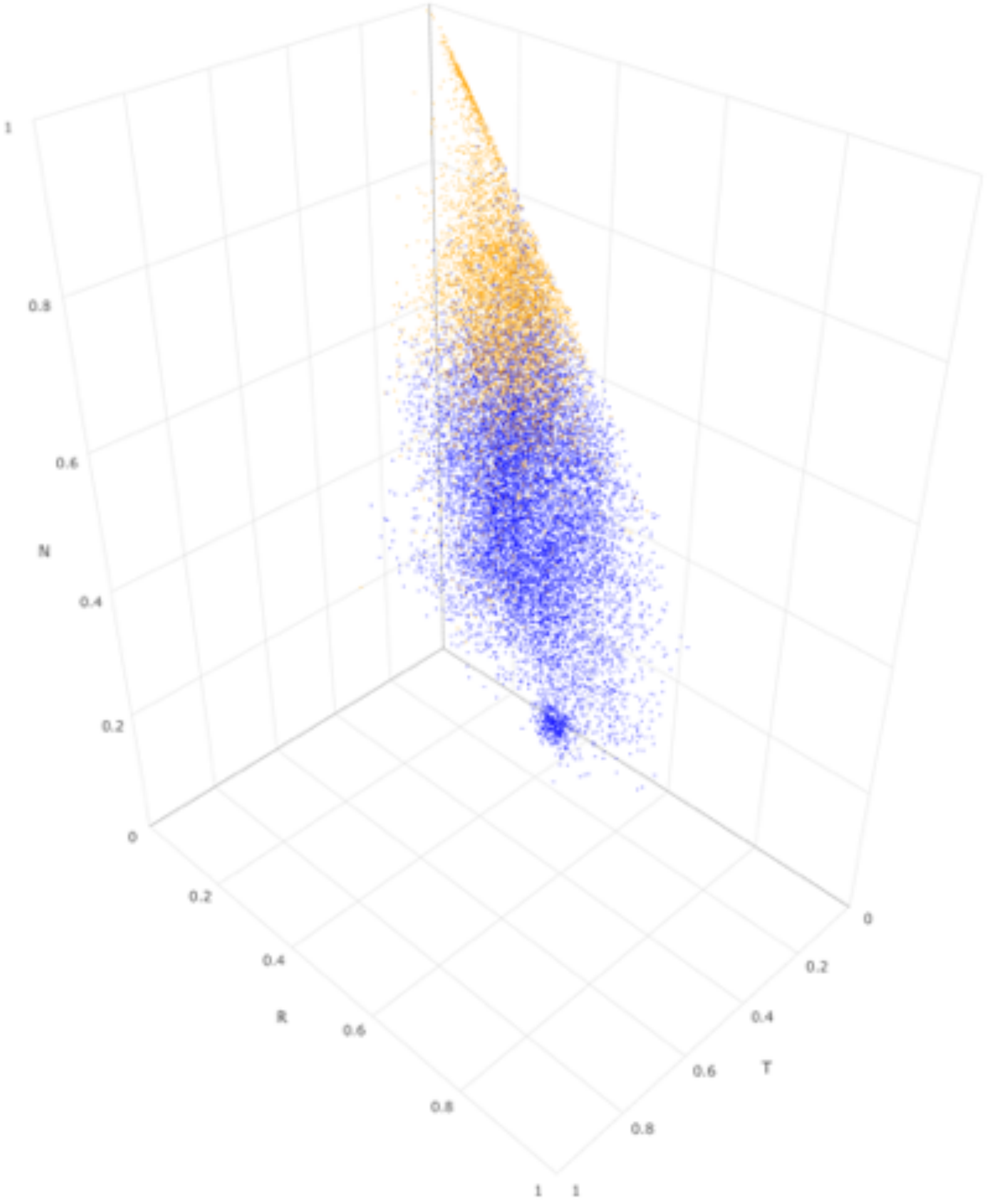

Fig. S11. *funtrp* prediction scores for disordered Proteins compared within position types. Proteins in Swiss-Prot were labeled as either ordered and disordered based on MetaDisorder predictions (Methods). Residues located in disordered proteins are highlited in yellow, those found in ordered proteins are shown in blue.

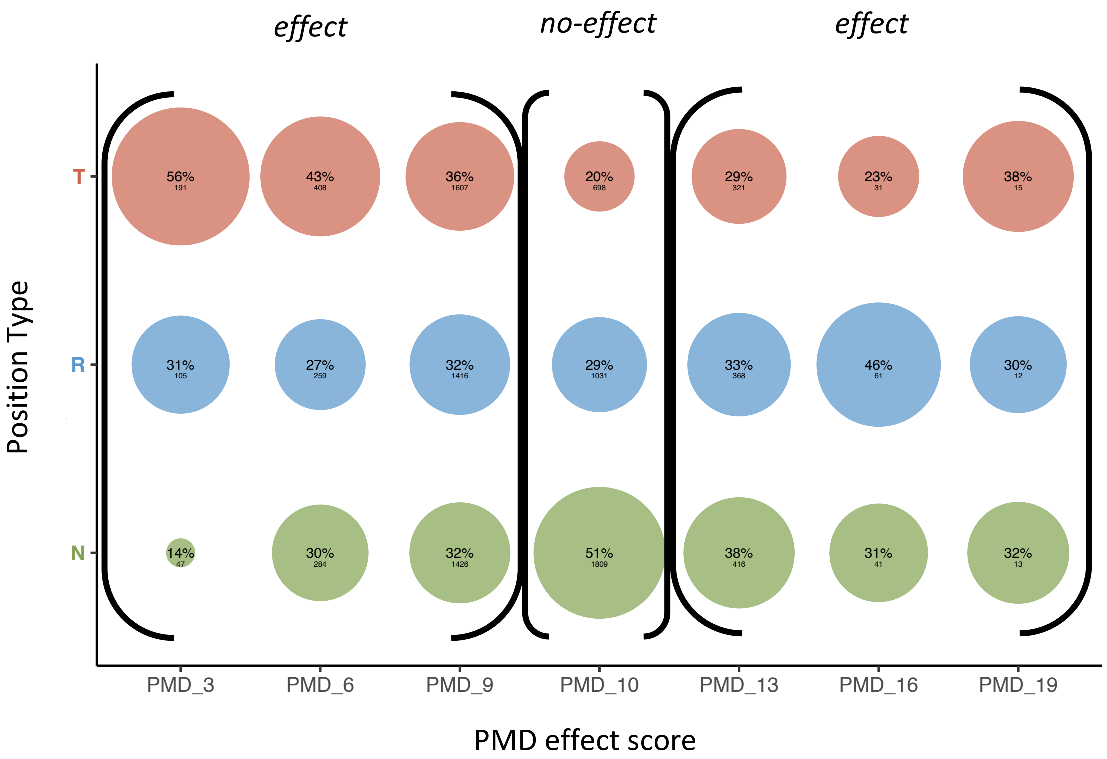

Fig. S12 Distribution of position types for PMD disease annotations. PMD score ranges (3-9: - / 10: = / 13-*19* +) were grouped into effect and no-effect. Percentages in (A) are rounded and thus do not add up to 100%.Colors are according to position type (green =Neutral, red =Rheostat, blue =Toggle).

Table S13. Performance of logistic regression models

method	Sensitivity	Specificity	Pos Pred Value	Neg Pred Value	Precision	Recall	F1	Balanced Accuracy	MCC
snap	0.660982	0.607573	0.627475	0.641861	0.627475	0.660982	0.643786	0.6342776	0.2689453
sift	0.554366	0.69174	0.642628	0.608211	0.642628	0.554366	0.5952343	0.6230528	0.2484609
pph2	0.691473	0.590406	0.627991	0.656816	0.627991	0.691473	0.6581998	0.6409395	0.2833394
funTRP	0.60711	0.647156	0.632447	0.622282	0.632447	0.60711	0.6194817	0.6271331	0.2544973
snap+funTRP	0.640329	0.656631	0.65094	0.646107	0.65094	0.640329	0.6455861	0.64848	0.2970034
sift+funTRP	0.614187	0.673562	0.652955	0.635821	0.652955	0.614187	0.6329718	0.6438745	0.2882617
pph2+funTRP	0.682427	0.623114	0.644223	0.662428	0.644223	0.682427	0.6627606	0.6527704	0.3060957

To compare *funtrp* with common variant effect prediction tools (SNAP, SIFT and PolyPhen-2) we converted predicted position types into approximated variant effect predictions (*Toggle* or *Rheostat* position = *effect* and *Neutral* = *no-effect*). We computed the performance for all four methods on the *no-effect vs. effect* groups extracted from PMD (described above). Performances are averaged over 100 iterations, each based on a subsampled dataset (without replacement and balanced regarding the class with fewer instances) from PMD.