Article

Disease variant prediction with deep generative models of evolutionary data

https://doi.org/10.1038/s41586-021-04043-8

Received: 18 December 2020

Accepted: 20 September 2021

Published online: 27 October 2011

Check for updates

Jonathan Frazer[1,4], Pascal Notin[2,4], Mafalda Dias[1,4], Aidan Gomez[2], Joseph K. Min[1], Kelly Brock[1], Yarin Gal[2]✉ & Debora S. Marks[1,3]✉

Quantifying the pathogenicity of protein variants in human disease-related genes would have a marked effect on clinical decisions, yet the overwhelming majority (over 98%) of these variants still have unknown consequences[1–3]. In principle, computational methods could support the large-scale interpretation of genetic variants. However, state-of-the-art methods[4–10] have relied on training machine learning models on known disease labels. As these labels are sparse, biased and of variable quality, the resulting models have been considered insufficiently reliable[11]. Here we propose an approach that leverages deep generative models to predict variant pathogenicity without relying on labels. By modelling the distribution of sequence variation across organisms, we implicitly capture constraints on the protein sequences that maintain fitness. Our model EVE (evolutionary model of variant effect) not only outperforms computational approaches that rely on labelled data but also performs on par with, if not better than, predictions from high-throughput experiments, which are increasingly used as evidence for variant classification[12–16]. We predict the pathogenicity of more than 36 million variants across 3,219 disease genes and provide evidence for the classification of more than 256,000 variants of unknown significance. Our work suggests that models of evolutionary information can provide valuable independent evidence for variant interpretation that will be widely useful in research and clinical settings.

The exponential growth in human genome sequencing has underlined the substantial genetic variation in the human population. Understanding the disease relevance of this genetic variation has the potential to transform healthcare and motivates the massive investment in the collection of human population genomic information together with demographics and clinical data such as the UK BioBank[1], ChinaMAP[17] and deCODE[18]. Access to sequencing has enabled both genetic studies that associate variants with diseases and more mechanism-based approaches that associate variants with biochemical and cellular phenotypes. However, relating specific changes in the genome to disease phenotypes remains an open challenge as the number of variants in the human population exceeds the number that we are able to investigate. Protein-coding regions alone contain large variation between people; to date, 6.5 million missense variants have been observed (gnomAD[2]) and the consequences of the vast majority (98%) of these, even in disease-related genes, are unknown[3]. It is estimated that there will be a variant for every protein position (bar lethal) somewhere in the human population of almost 8 billion.

Given this challenge, new experimental technologies have emerged that can assess the effects of thousands of mutations in parallel (sometimes called deep mutational scans or multiplexed assays of variant effects (MAVEs))[19,20]. The results of these high-throughput experiments are then scrutinized by expert panels such as ClinGen[21] for assigning clinical interpretation to human variants. However, these technologies do not easily scale to thousands of proteins, especially not to combinations of variants, and depend critically on the availability of assays that are relevant to or at least associated with human disease phenotypes.

Ideally, computation could also accelerate clinical variant interpretation. However, state-of-the-art computational methods[4–10] are supervised on clinical labels in a way that causes inflated accuracy in real-world prediction scenarios. This inflated performance results from variant aggregation across genes (label bias), label sparsity, label noise[12] and data leakage[22] (Supplementary Note 1). It is hardly surprising, therefore, that there is some hesitation to use computational methods for anything but 'weak evidence' for variant classification, as in the guidelines from the American College of Medical Genetics and Association for Molecular Pathology[11]. By contrast, unsupervised probabilistic models of evolutionary sequences alone have been notably successful at predicting the effects of variants on protein function and stability[23–26] and are fundamentally generalizable as they avoid learning from labels. However, there has been little progress in developing these models to address disease relevance since early pioneering efforts[27,28].

In this work, we revisit the clinical value of evolutionary information in light of recent developments in unsupervised generative modelling. We introduce EVE, a computational method for the classification of human genetic variants trained solely on evolutionary sequences. We show that EVE outperforms current state-of-the-art computational methods at predicting variant pathogenicity (without the risk

[1]Marks Group, Department of Systems Biology, Harvard Medical School, Boston, MA, USA. [2]OATML Group, Department of Computer Science, University of Oxford, Oxford, UK. [3]Broad Institute of Harvard and MIT, Cambridge, MA, USA. [4]These authors contributed equally: Jonathan Frazer, Pascal Notin, Mafalda Dias. ✉e-mail: yarin.gal@cs.ox.ac.uk; debbie@hms.harvard.edu

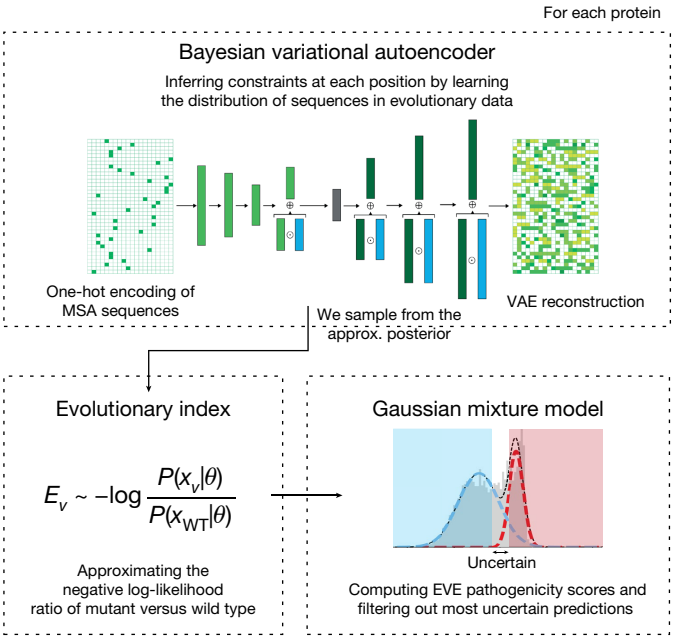


Fig. 1 | Modelling strategy. For each protein, a Bayesian VAE (top) learns a distribution over amino acid sequences in a multiple sequence alignment (MSA) of evolutionary data. This enables the computation of the evolutionary index (bottom left) for each single-variant sequence, which approximates the negative log-likelihood ratio of variant (X_v) versus wild-type (X_{WT}) sequences. A global-local mixture of a Gaussian mixture model (bottom right) separates variants into benign (blue dashed line) and pathogenic (red) clusters based on that index. The outcome of the model is both a continuous score that reflects pathogenicity propensity, and probabilistic assignment to benign and pathogenic classes (blue and red shaded areas, respectively) below a user-defined uncertainty threshold (Extended Data Figs. 1, 3).

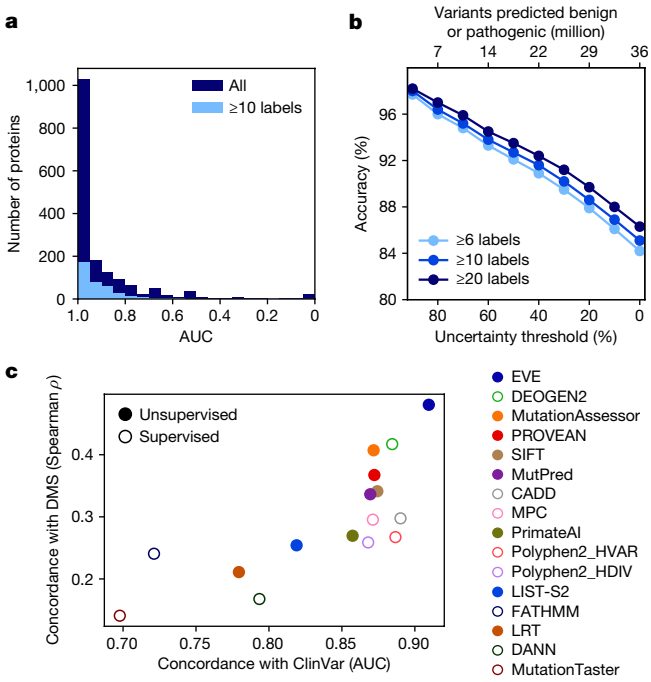


Fig. 2 | EVE accurately predicts disease-causing variants. a, Distribution of the AUC for EVE scores computed over known clinical labels from ClinVar, on all 3,219 proteins covered by our study (dark blue) and for the subset of proteins with at least five benign and five pathogenic known labels (pale blue). **b**, Trade-off between the accuracy of EVE and the uncertainty threshold (percentage of variants set as 'uncertain') or the total number of variants given a class assignment. Accuracy was computed over all labels for proteins with at least three (five or ten) benign labels and three (five or ten) pathogenic labels. **c**, Performance comparison of EVE to state-of-the-art computational variant effect predictors: seven unsupervised and eight supervised. Performance was estimated against known clinical labels (average AUC over disease genes in ClinVar (*x* axis)), and against high-throughput functional assays developed to assess the clinical effect of variants (average spearman correlation (*y* axis)) (Supplementary Note 2, Extended Data Fig. 4, Supplementary Tables 2–4). DMS, deep mutational scan.

of overfitting clinical labels) and is as accurate as predictions from high-throughput experiments.

Predicting pathogenicity from evolution

Our method—EVE—learns the propensity of human missense variants to be pathogenic from the distribution of sequence variation across species (Fig. 1, Extended Data Fig. 1). In the first step, we captured constraints from natural sequences across evolution, including complex dependencies between positions, by learning the distribution of amino acid sequences for each protein using an expressive deep generative model, a variational autoencoder (VAE)[29,30]. VAEs have been successful in learning complex high-dimensional distributions across multiple domains including prediction of protein function[31] (Supplementary Methods). For each human protein of interest, a Bayesian VAE was trained on a multiple sequence alignment retrieved by searching approximately 250 million protein sequences in UniRef[32] (Supplementary Methods, Supplementary Table 1). After training on evolutionary sequences, we estimated the relative likelihood of each single amino acid variant with respect to the wild type—what we call the 'evolutionary index'—by sampling from the approximate posterior distribution learned by the VAE. We performed a thorough architecture and hyperparameter search to ensure stability and performance across proteins, and demonstrate its superiority over previous methods[31] (Extended Data Fig. 2). When comparing this evolutionary index against clinical labels, the value that separates pathogenic from benign labels was notably consistent across proteins (Extended Data Fig. 3a), suggesting that we may use unsupervised methods to infer pathogenicity. Therefore, in the second step, rather than using (semi-)supervised learning to map scores to label categories, we fitted a two-component global-local mixture of Gaussian mixture models on the distributions of evolutionary indices for all single amino acid variants across proteins (Extended Data Fig. 3b, Supplementary Methods). The output of this process is both the EVE score—a continuous pathogenicity score defined over the interval [0,1], with zero being most benign and one being most pathogenic—and class assignments. For these assignments, we used the predictive entropy of the Gaussian mixture model as a measure of classification uncertainty, and binned variants into one of three categories: benign, uncertain or pathogenic (Supplementary Methods).

We applied EVE to a set of 3,219 human genes that have been associated with disease in ClinVar[3] (Supplementary Methods). Our model is predictive of clinical significance for all labelled variants across all genes (average area under the curve (AUC) of 0.91) (Fig. 2b, Supplementary Table 2) including 60 'clinically actionable' genes[33] (average AUC of 0.92) (Extended Data Fig. 4a). Furthermore, the performance of EVE is robust to the number of labels per protein (Fig. 2b), suggesting generalizability to genes with less (or no) annotation, as we would expect from an unsupervised approach.

EVE outperforms all supervised and unsupervised methods at predicting known clinical labels (Fig. 2c, *x* axis, Supplementary Table 3). This is despite a large fraction of these labels being used in training the top-performing methods, as well as, in some cases, being used extensively in defining labels. As a second benchmark that avoids some of these circularities, we compared the model predictions against 40,000 experimentally measured variants across 10 proteins

(Supplementary Methods). As these experiments are, in principle, independent of the ClinVar labelling process, we expected this benchmark to provide a less biased estimate of performance, albeit for a comparatively smaller number of proteins. On this benchmark, EVE outperforms all methods (Fig. 2c, *y* axis, Supplementary Table 4), including meta-predictors (Extended Data Fig. 4b, Supplementary Note 2, Supplementary Methods).

As the consequence of variant classification significantly varies from one gene to another, an important feature of our method is the ability to assign a degree of uncertainty to the prediction, allowing a trade-off between predicted accuracy and coverage of variants. Setting aside an increasing number of variants as 'uncertain' enabled us to reach higher accuracy over the variants that we do classify as pathogenic or benign. For instance, excluding the 25% of most uncertain variants resulted in an accuracy of approximately 90% for pathogenic and benign classifications (Fig. 2b, Supplementary Table 2). In practice, we envision researchers deciding on specific trade-offs on a gene-by-gene and use case basis.

EVE is as accurate as experimental prediction

We asked whether our computational predictions are as accurate as experimental predictions. For the five genes with a large number of high-quality labels in ClinVar (*BRCA1*, *TP53*, *PTEN*, *MSH2* and *SCN5A*), the overall performance of EVE at predicting clinical significance is as good as, or better than, that of the deep mutational scan experiments that were specifically designed to predict pathogenicity[12–16] (Fig. 3, Extended Data Fig. 5, Supplementary Methods). For instance, for *TP53*, EVE predicted near-perfect separation of benign and pathogenic variants for the whole protein in contrast to the experimental predictions[14] that are weaker in the tetramer domain (from position 300 to the end). For *SCN5A* (which is associated with Brugada syndrome[34] and long QT syndrome[35]), EVE predicted R814Q to be pathogenic even though this is a gain, rather than loss of function, in the experiments from ref. [36], suggesting that evolutionary data contain information about gain of function and support the known genetics[35]. EVE also has marginally better performance than experiments on a larger set of genes that have fewer high-quality labels (Extended Data Fig. 6, Supplementary Methods, Supplementary Tables 5, 6).

As EVE and MAVEs are independent sources of evidence, comparison of their results may help to evaluate the clinical labels themselves. Across *MSH2*, *PTEN* and *TP53*, 23 of the 27 variants (85%) where the EVE score disagrees with ClinVar, MAVE experimental data support the EVE classification. Both EVE and experiments support a benign score for variants R337H and R337C in *TP53*, and S554N/T, D660G and I774V in *MSH2*, and 15 variants in the *PTEN* score where ClinVar has pathogenic labels. Similarly, both EVE and experimental assays support a pathogenic clinical effect where ClinVar has benign labels for G759E and E198G in *MSH2* (the pathogenic assignment of the latter is further supported by new experimental data[37]). An obvious caveat where concordance between functional assay prediction and EVE may be misleading is the case of functional RNA, for example, splice variation[16].

Together, our analysis shows that EVE prediction performs as well as predictions from high-throughput experiments, suggesting that it may be beneficial to focus experimental efforts on genes where EVE does not perform well (Supplementary Table 7).

Predictions for 36 million variants

We provide both continuous EVE scores and class assignments for the 36 million single amino acid variants across the 3,219 disease-associated genes. Of these variants, approximately 1.3 million have been observed in at least one human to date (UK Biobank[1] and gnomAD[2]; Supplementary Methods), but only about 3% have any clinical interpretation in ClinVar (Fig. 4a, left). The EVE class assignments, after dropping the 25% most uncertain variants to keep accuracy at approximately 90%,

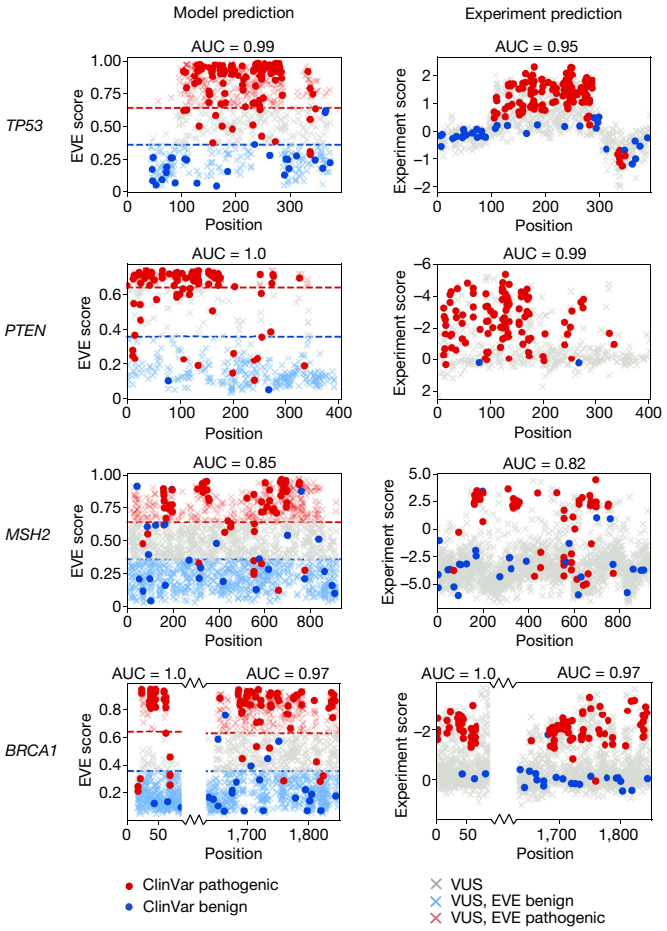


Fig. 3 | EVE is as good as functional experiments at predicting clinical interpretations of variants. Comparison of computational model predictions (left panels, *y* axis EVE score) and experimental predictions (right panels, *y* axis experimental score) to ClinVar labels (dots) and variants of unknown significance (VUS) (crosses), where pale red and pale blue crosses indicate EVE predictions; the *x* axis corresponds to the position in the protein. The dashed red and blue lines correspond to EVE predictions setting the 25% most uncertain assignments as uncertain (Supplementary Methods). Experimental data from deep mutational scans of *TP53* (ref. [13]), *PTEN*[19], *MSH2* (ref. [43]) and *BRCA1* (ref. [12]) (Extended Data Fig. 5, Supplementary Table 6).

provide an interpretation for about 27 million variants in total and over 800,000 (approximately 64%) of the variants seen to date in humans (Fig. 4a, middle, Supplementary Methods).

The continuous scores for all single amino acid variants provide a complementary picture to that of class assignments. The distribution of EVE scores within proteins highlights clusters of high pathogenicity, following trends that might be expected by functional importance, such as hydrophobic cores, and ligand-binding and active sites. For instance, many variants with high EVE scores in the SCN4A–SCN1B ion channel complex (PDB 6AGF[38]) lie at the complex interface, line the SCN4A pore and the hydrophobic core of SCN1B (Fig. 4b, c). For the mismatch DNA repair complex MSH2–MSH6 (associated with Lynch syndrome[39] and approximately 20% of sporadic cancers[40]), EVE pathogenic signals are strong for variants that are proximal to the bound ADP and DNA (PDB 2O8B[41]) where clinical labels are sparse (yet observed in the population) (Fig. 4d).

Combining EVE with other evidence

EVE provides a single source of evidence, making it ideal for combining with other, orthogonal sources of evidence (as is typically performed by

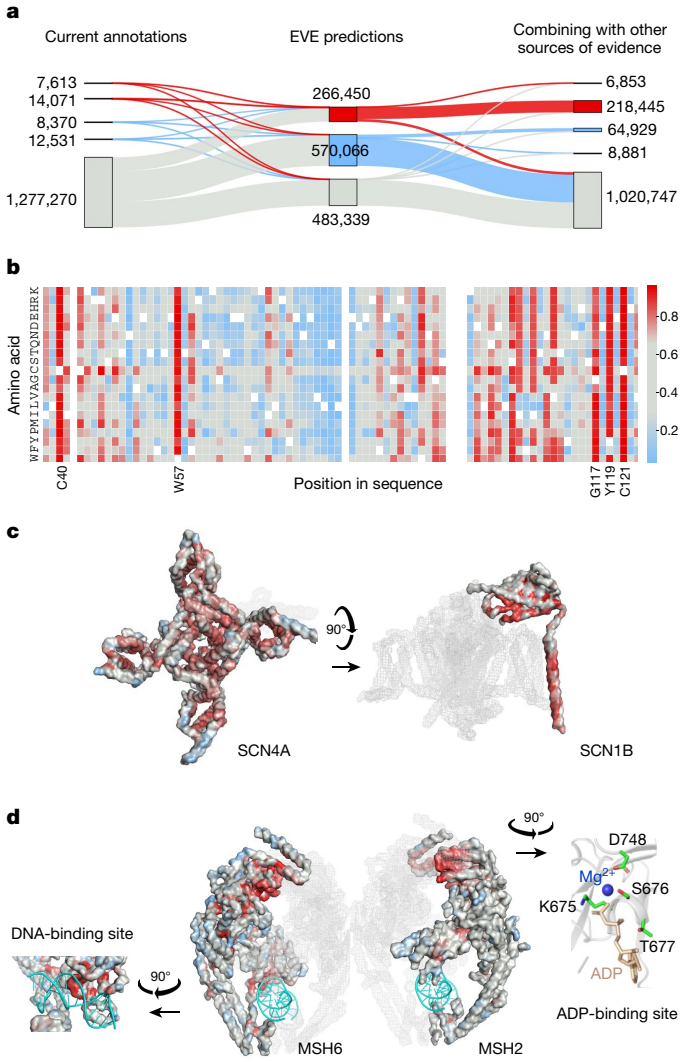


Fig. 4 | Predictions for variants in 3,219 genes. a, Combining EVE classifications with other sources of evidence. ClinVar labels and VUS based on gnomAD and UK Biobank (left); EVE predictions setting 25% of all possible variants as uncertain (middle); and predictions after combining EVE with other sources of evidence (right) (Supplementary Methods; Supplementary Table 2). **b**, Heat map of EVE pathogenicity scores in SCN1B. **c**, **d**, Representations of 3D structures of SCN4A–SCN1B (PDB 6AGF[38]) (**c**) and MSH6–MSH2 bound to ADP and a G T mispair (PDB 2O8B[41]) (**d**), coloured by mean score per position (SCN4A, MSH6 and MSH2) and maximum score per position (SCN1B). Clusters of high pathogenicity in 3D include the pore region of SCN4A, the hydrophobic core of SCN1B (positions 40, 57, 117, 119 and 121), the C terminus α-helix of SCN1B and the interface with SCN4A, the ADP-binding site of MSH2 (such as D748N/V/H, K675E, S676L and T677R) and the DNA-binding site of MSH6.

expert panels, for example, ClinGen[21]). To illustrate this, we combined our model class assignments with population data from gnomAD[2] and other forms of existing evidence. This resulted in 256,000 variants with no previous clinical interpretation for potential reclassification, and another 539 variants that contradict current ClinVar status for which we found independent supporting evidence. Examples of the latter include the *MSH2* variants described above, and the *TP53* variant R337Q (Fig. 4a, Supplementary Table 8, Supplementary Methods).

Being unsupervised also opens the door to a more refined approach in which the strength of evidence provided by the model may be allowed to vary on a gene-by-gene basis, in close analogy with recommendations for functional assays[42]. This offers a clear advantage over supervised methods. For example, if we consider 1,000 genes with at least 10 labels for validation, a supervised method (using a 90% train, 10% test and random split) leaves only approximately 50 proteins on which to test (Extended Data Fig. 7).

Discussion

It has long been appreciated that looking at the patterns of sequence conservation across species can yield insights into the consequences of variation within a species[43,44], including insights into human variants and disease association[45]. By bringing together recent developments in machine learning with the rapidly increasing amount of sequencing data from diverse organisms, we can extract more precise statements than previously realized and on a sufficiently large scale to be able to impact our sum knowledge of the clinical significance of variants. All data, results and code are available at or linked from https://evemodel.org/ (Extended Data Fig. 8), which will be regularly updated with new genes.

We have demonstrated that deep generative models trained on sequence alignments alone achieve state-of-the-art performance on variant classification and do so while avoiding the issues that typically affect supervised methods. This not only leads to better generalization guarantees but also provides a source of evidence that is independent and complementary to other large-scale efforts (for example, population data from biobanks) and yields an order of magnitude gain of scope when validating on a gene-by-gene basis. Although we do not know precisely how the constraints learnt from the sequences relate to disease, we observed a performance on par with functional assays in predicting pathogenicity. This suggests that expert panels could subject our method to similar scrutiny for classification as experiments like MAVEs.

The primary advantage of our approach over experimental approaches is significant gain in scope at a negligible fraction of the cost. An appealing prospect is that our method may be useful in guiding future experimental efforts, essentially acting as a means of identifying which variants and which genes would be most informative to probe (Supplementary Table 7).

There are important challenges in assessing missense variants that are not covered in this report. First, is the heterogeneity of disease presentation; we know that different variants of the same gene, and even the same variant, can lead to different disease severity or even different diseases, aspects that will be masked by the use of simple discrete pathology categories. Although we expect that the continuous EVE score, as opposed to the discrete classifications, may be useful to predict disease severity, this remains speculative and does not account for entirely different disease presentations from variants in the same gene. Second, this current work does not explicitly address the effect of combinations of variants. As humans have on average approximately 12% of their genes with two or more variants compared with a reference genome, albeit not necessarily on the same chromosome, this will be an important consideration. For the 'ACMG actionable genes', there are approximately 21,000 distinct pairs of variants in the same gene, occurring 1.5 million times (UK BioBank[1]; Extended Data Fig. 9, Supplementary Methods, Supplementary Table 9).

We conclude with a remark on biodiversity. Our analysis is one small but unusually direct demonstration of how the diversity of life on Earth benefits human health. Our models make use of data from over 140,000 organisms. Of these, we identified 17,000 organisms that are on the International Union for Conservation of Nature's Red List of Threatened Species[46] including 1,301 classified as vulnerable, 1,148 endangered, 548 critically endangered, 10 extinct in the wild and 21 extinct organisms. The progressive disappearance of species is a threat to the diversity on which this work is built.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information,

acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-021-04043-8.

1. Van Hout, C. V. et al. Exome sequencing and characterization of 49,960 individuals in the UK Biobank. *Nature* **586**, 749–756 (2020).
2. Karczewski, K. J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
3. Landrum, M. J. & Kattman, B. L. ClinVar at five years: delivering on the promise. *Hum. Mutat.* **39**, 1623–1630 (2018).
4. Raimondi, D. et al. DEOGEN2: prediction and interactive visualization of single amino acid variant deleteriousness in human proteins. *Nucleic Acids Res.* **45**, W201-W206 (2017).
5. Feng, B. J. PERCH: a unified framework for disease gene prioritization. *Hum. Mutat.* **38**, 243–251 (2017).
6. Ioannidis, N. M. et al. REVEL: an ensemble method for predicting the pathogenicity of rare missense variants. *Am. J. Hum. Genet.* **99**, 877-885 (2016).
7. Ionita-Laza, I., McCallum, K., Xu, B. & Buxbaum, J. D. A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nat. Genet.* **48**, 214–220 (2016).
8. Jagadeesh, K. A. et al. M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. *Nat. Genet.* **48**, 1581-1586 (2016).
9. Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* **47**, D886–D894 (2019).
10. Adzhubei, I. A. et al. A method and server for predicting damaging missense mutations. *Nat. Methods* **7**, 248–249 (2010).
11. Richards, S. et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.* **17**, 405–424 (2015).
12. Findlay, G. M. et al. Accurate classification of BRCA1 variants with saturation genome editing. *Nature* **562**, 217–222 (2018).
13. Glazer, A. M. et al. High-throughput reclassification of SCN5A variants. *Am. J. Hum. Genet.* **107**, 111–123 (2020).
14. Giacomelli, A. O. et al. Mutational processes shape the landscape of TP53 mutations in human cancer. *Nat. Genet.* **50**, 1381–1387 (2018).
15. Mighell, T. L., Evans-Dutson, S. & O'Roak, B. J. A saturation mutagenesis approach to understanding PTEN lipid phosphatase activity and genotype–phenotype relationships. *Am. J. Hum. Genet.* **102**, 943–955 (2018).
16. Jia, X. et al. Massively parallel functional testing of MSH2 missense variants conferring Lynch syndrome risk. *Am. J. Hum. Genet.* **108**, 163–175 (2021).
17. Cao, Y. et al. The ChinaMAP analytics of deep whole genome sequences in 10,588 individuals. *Cell Res.* **30**, 717–731 (2020).
18. Gudbjartsson, D. F. et al. Large-scale whole-genome sequencing of the Icelandic population. *Nat. Genet.* **47**, 435–444 (2015).
19. Esposito, D. et al. MaveDB: an open-source platform to distribute and interpret data from multiplexed assays of variant effect. *Genome Biol.* **20**, 223 (2019).
20. Trenkmann, M. Putting genetic variants to a fitness test. *Nat. Rev. Genet.* **19**, 667 (2018).
21. Rehm, H. L. et al. ClinGen—the Clinical Genome Resource. *N. Engl. J. Med.* **372**, 2235–2242 (2015).
22. Grimm, D. G. et al. The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity. *Hum. Mutat.* **36**, 513–523 (2015).
23. Hopf, T. A. et al. Mutation effects predicted from sequence co-variation. *Nat. Biotechnol.* **35**, 128–135 (2017).
24. Marks, D. S. et al. Protein 3D structure computed from evolutionary sequence variation. *PLoS ONE* **6**, e28766 (2011).
25. Hopf, T. A. et al. Sequence co-evolution gives 3D contacts and structures of protein complexes. *eLife* **3**, e03430 (2014).
26. Lapedes, A., Giraud, B. & Jarzynski, C. Using sequence alignments to predict protein structure and stability with high accuracy. Preprint at https://arxiv.org/abs/1207.2484v1 (2012).
27. Vaser, R., Adusumalli, S., Leng, S. N., Sikic, M. & Ng, P. C. SIFT missense predictions for genomes. *Nat. Protoc.* **11**, 1–9 (2016).
28. Reva, B., Antipin, Y. & Sander, C. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res.* **39**, e118 (2011).
29. Rezende, D. J., Mohamed, S. & Wierstra, D. in *Proceedings of the 31st International Conference on Machine Learning* vol. 32 (eds Xing, E. P. & Jebara, T.) 1278–1286 (PMLR, 2014).
30. Kingma, D. P. & Welling, M. Auto-encoding variational Bayes. Preprint at https://arxiv.org/abs/1312.6114 (2013).
31. Riesselman, A. J., Ingraham, J. B. & Marks, D. S. Deep generative models of genetic variation capture the effects of mutations. *Nat. Methods* **15**, 816–822 (2018).
32. Suzek, B. E., Wang, Y., Huang, H., McGarvey, P. B. & Wu, C. H. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* **31**, 926–932 (2015).
33. Kalia, S. S. et al. Recommendations for reporting of secondary findings in clinical exome and genome sequencing, 2016 update (ACMG SF v2.0): a policy statement of the American College of Medical Genetics and Genomics. *Genet. Med.* **19**, 249–255 (2017).
34. Frigo, G. et al. Homozygous SCN5A mutation in Brugada syndrome with monomorphic ventricular tachycardia and structural heart abnormalities. *Europace* **9**, 391–397 (2007).
35. Itoh, H. et al. Asymmetry of parental origin in long QT syndrome: preferential maternal transmission of KCNQ1 variants linked to channel dysfunction. *Eur. J. Hum. Genet.* **24**, 1160–1166 (2016).
36. Glazer, A. M. et al. Deep mutational scan of an SCN5A voltage sensor. *Circ. Genom. Precis. Med.* **13**, e002786 (2020).
37. Bouvet, D. et al. Methylation tolerance-based functional assay to assess variants of unknown significance in the MLH1 and MSH2 genes and identify patients with Lynch syndrome. *Gastroenterology* **157**, 421–431 (2019).
38. Pan, X. et al. Structure of the human voltage-gated sodium channel $Na_v1.4$ in complex with β1. *Science* **362**, eaau2486 (2018).
39. Fishel, R. et al. The human mutator gene homolog MSH2 and its association with hereditary nonpolyposis colon cancer. *Cell* **75**, 1027–1038 (1993).
40. Peltomaki, P. Role of DNA mismatch repair defects in the pathogenesis of human cancer. *J. Clin. Oncol.* **21**, 1174-1179 (2003).
41. Warren, J. J. et al. Structure of the human MutSα DNA lesion recognition complex. *Mol. Cell* **26**, 579–592 (2007).
42. Brnich, S. E. et al. Recommendations for application of the functional evidence PS3/BS3 criterion using the ACMG/AMP sequence variant interpretation framework. *Genome Med.* **12**, 3 (2019).
43. Lewontin, R. C. *The Genetic Basis of Evolutionary Change* (Columbia Univ. Press, 1974).
44. Kreitman, M. Nucleotide polymorphism at the alcohol dehydrogenase locus of *Drosophila melanogaster*. *Nature* **304**, 412-417 (1983).
45. Sunyaev, S. et al. Prediction of deleterious human alleles. *Hum. Mol. Genet.* **10**, 591–597 (2001).
46. IUCN. The IUCN red list of threatened species. *IUCN* https://www.iucnredlist.org (2020).

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

The data analysed and generated in this study, including multiple sequence alignments used in training, ClinVar annotations used for validation, population frequencies and predictions from our model, are available in Supplementary Information and at evemodel.org. Predictions from other computational models are available through http://database.liulab.science/dbNSFP. Source data are provided with this paper.

Code availability

The model code is available at https://github.com/OATML-Markslab/EVE, https://doi.org/10.5281/zenodo.5389490.

Acknowledgements We thank members of the Marks laboratory, OATML and C. Sander for many valuable discussions. J.F., M.D. and K.B. are supported by the Chan Zuckerberg Initiative CZI2018-191853. K.B. is also supported by the US National Institutes of Health (R01 R01GM120574). P.N. is supported by GSK and the UK Engineering and Physical Sciences Research Council (EPSRC ICASE award no. 18000077). A.G. is a Clarendon Scholar and Open Philanthropy AI Fellow. Y.G. holds a Turing AI Fellowship (Phase 1) at the Alan Turing Institute, which is supported by EPSRC grant reference V030302/1. D.S.M. holds a Ben Barres Early Career Award by the Chan Zuckerberg Initiative as part of the Neurodegeneration Challenge Network, CZI2018-191853.

Author contributions D.S.M. and Y.G. led the research. J.F., P.N. and M.D. conceived and implemented the end-to-end approach. A.G. contributed technical advice. K.B. supported with data preparation. J.K.M. developed the website. J.F., P.N., M.D., Y.G. and D.S.M. wrote the manuscript.

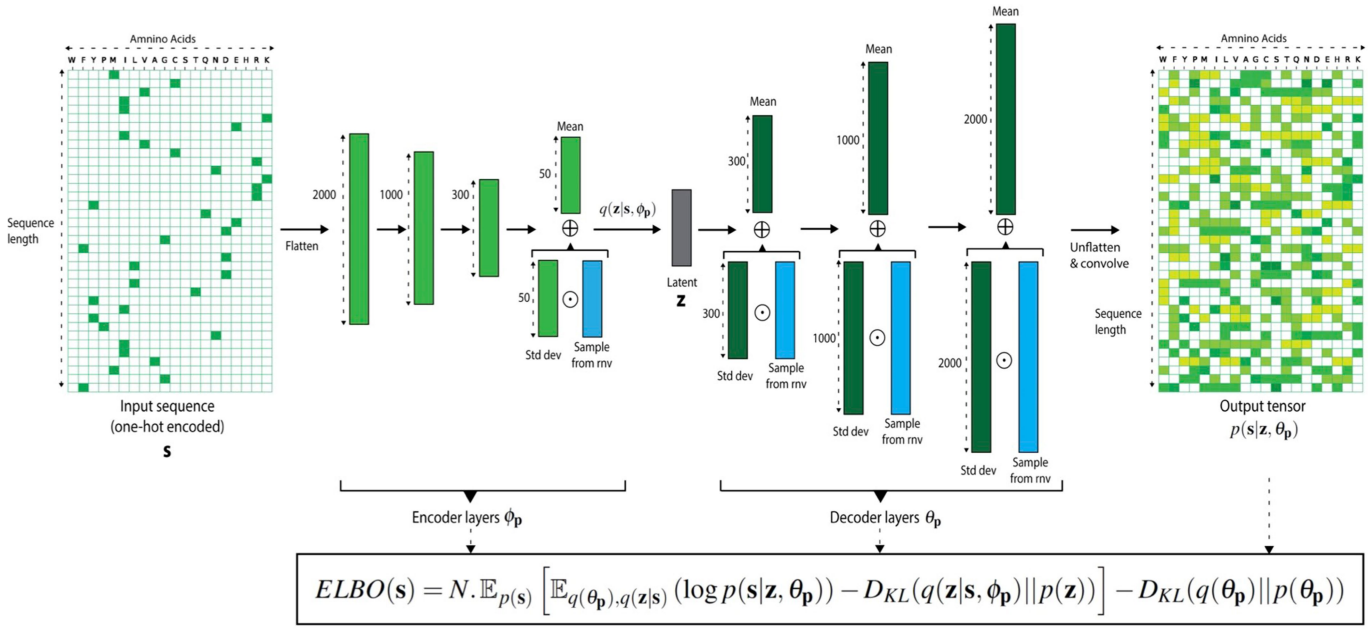
Competing interests The authors declare no competing interests.

Additional information

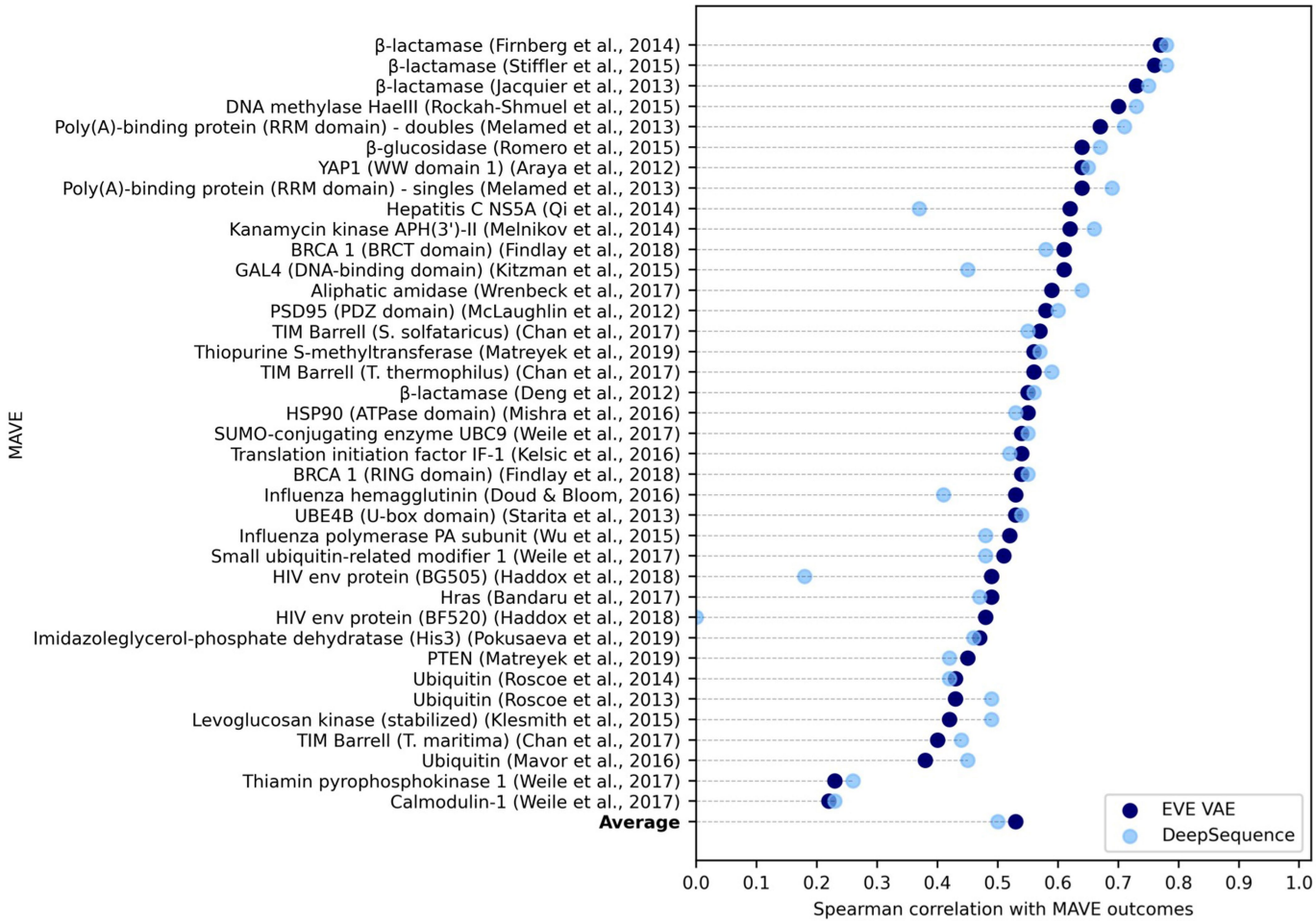
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-021-04043-8.

Correspondence and requests for materials should be addressed to Yarin Gal or Debora S. Marks.

Peer review information *Nature* thanks Martin Kircher and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

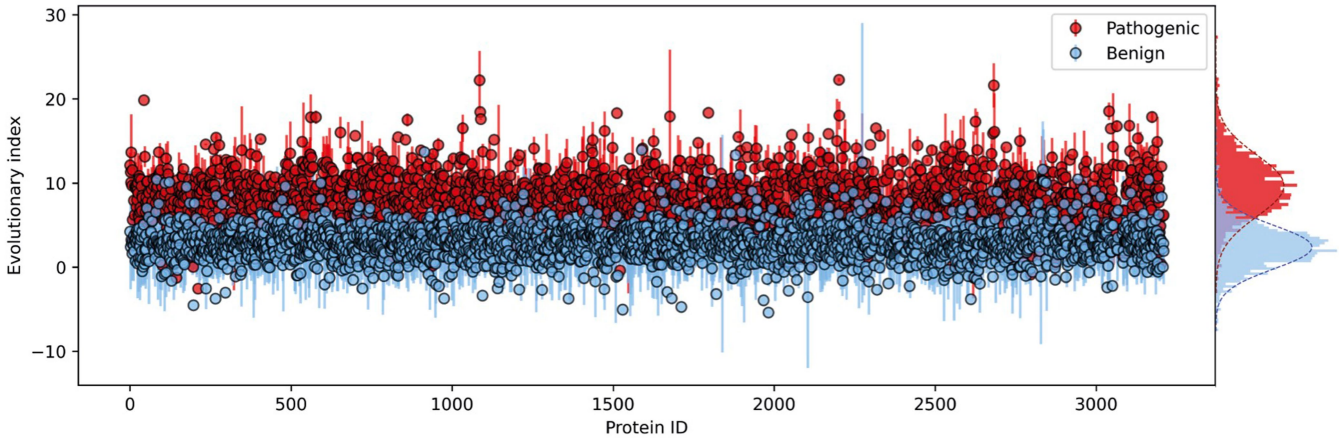


Extended Data Fig. 1 | Bayesian VAE architecture details. The Bayesian VAE architecture in EVE is comprised of a symmetric 3-layer encoder & decoder architecture (with 2,000-1,000-300 and 300-1,000-2,000 units respectively) and a latent space of dimension 50. After performing a one-hot encoding of the input sequence across amino acids (zeros in white, ones in green), we flatten the input before performing the forward pass through the network. We use a single set of parameters for the encoder (ϕp) and learn a fully-factorized gaussian distribution over the weights of the decoder (θp): weight samples for the decoder are obtained by sampling a random normal variable (rnv), multiplying that sample by the standard deviation parameters, and subsequently adding the mean parameters. A one-dimensional convolution is applied on the un-flattened output of the decoder to capture potential correlations between amino-acid usage. Finally, a softmax activation turns the final output into probabilities over amino acids at each position of the sequence (low values in white, high values in dark green). The overall network is trained by maximizing the Evidence Lower Bound (ELBO), which forms a tractable lower bound to the log-marginal likelihood (Supplementary Methods and Fig. 1).

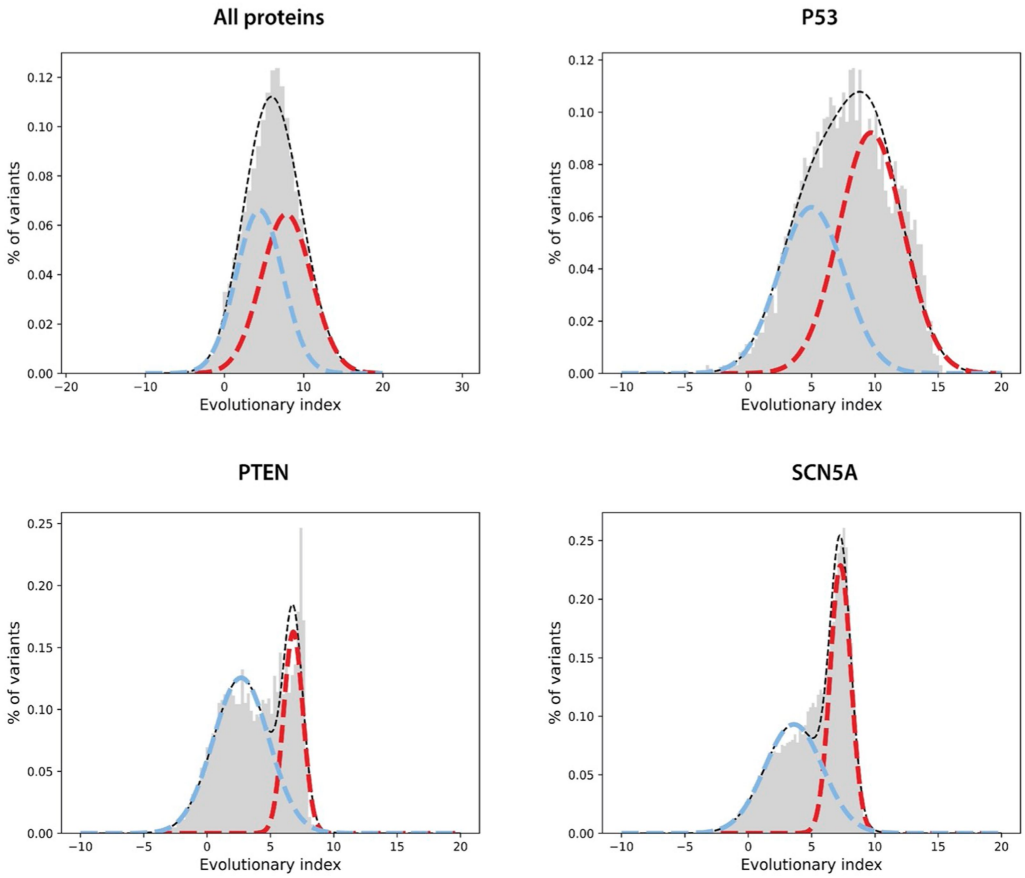


Extended Data Fig. 2 | Comparison of performance of Bayesian VAE and DeepSequence against 38 deep mutation scans. Comparison between the performance of the Bayesian VAE architecture in EVE and DeepSequence[46] which achieves state-of-the-art performance on the protein function prediction task. "Evolutionary indices" were computed by sampling 2k times from the approximate posterior distribution and by ensembling the obtained indices over 5 independently trained VAEs (Supplementary Methods).

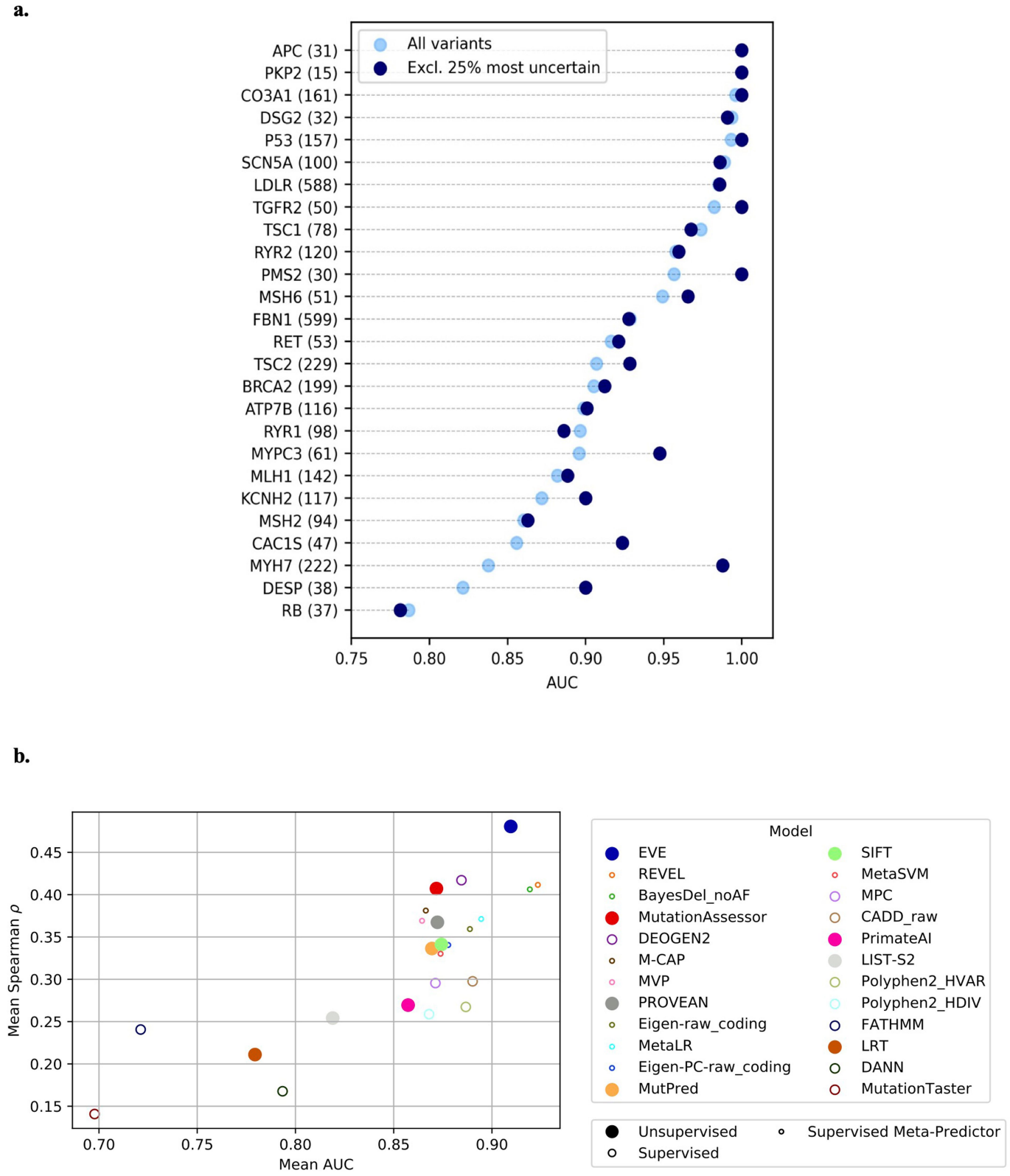
a.



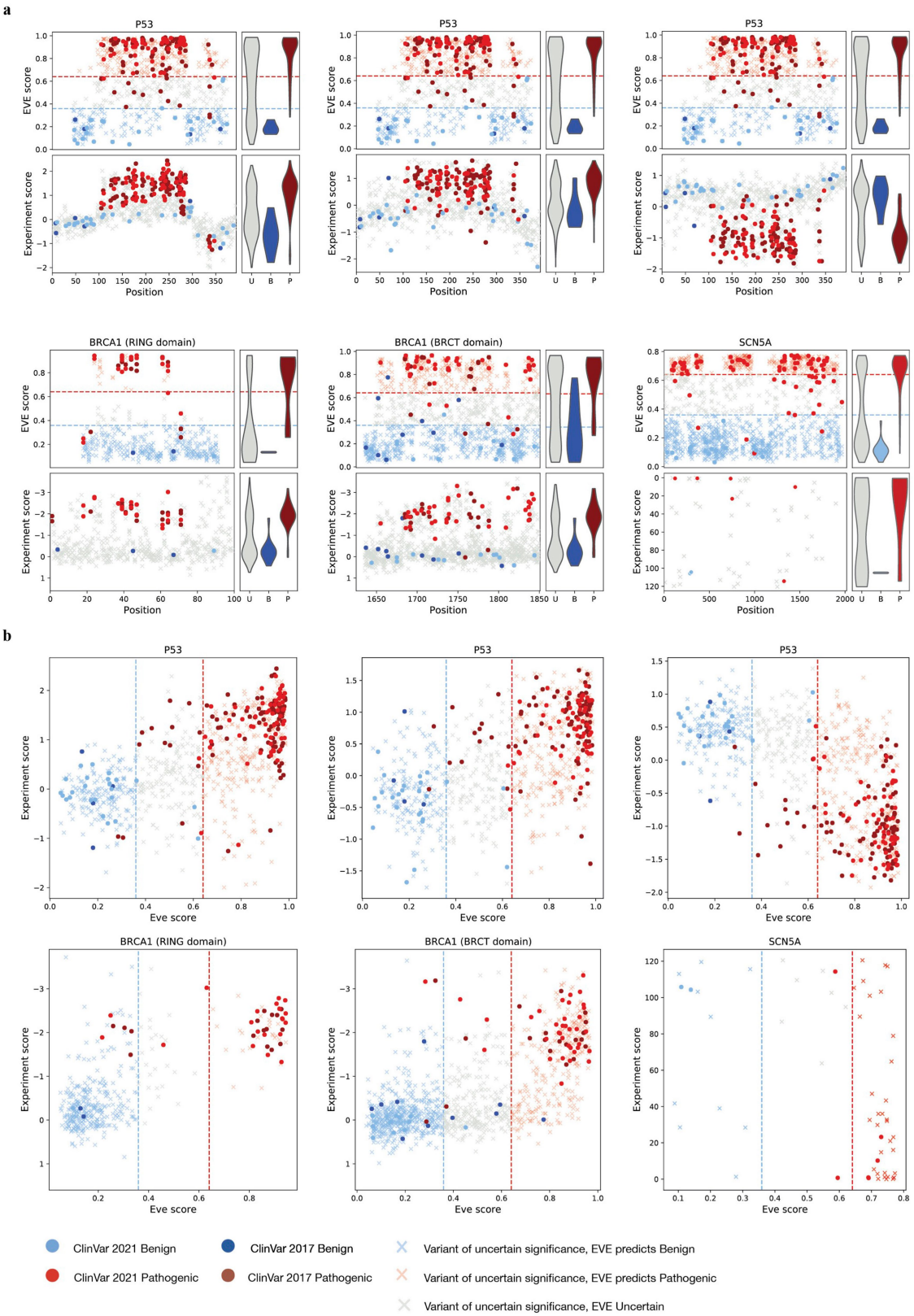
b.



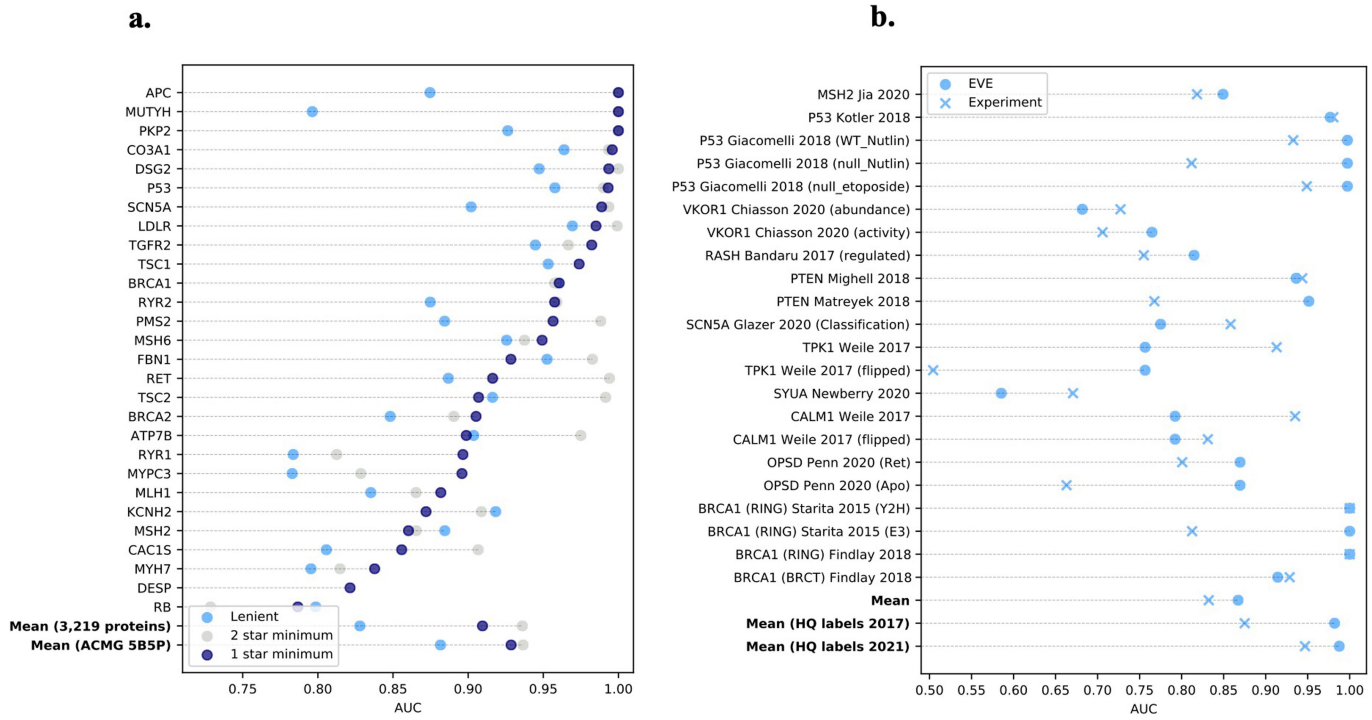
Extended Data Fig. 3 | Evolutionary index separates pathogenic and benign variants. a, Average evolutionary index per protein, and corresponding standard deviations, for variants with known Benign and Pathogenic ClinVar labels across 3,219 proteins (sorted by alphabetical order). On the right, marginal distributions of the means over the 3,219 proteins. Evolutionary index separates pathogenic and benign labels consistently across proteins. **b**, Two-component Gaussian Mixture Models (GMM) over the distributions of the evolutionary indices (histograms) for all the single amino acid variants of 3,219 proteins combined (top, left) and for P53, PTEN and SCN5A separately (top right, bottom left and right, respectively). The dashed black line is the marginal likelihood for the GMM model, i.e. the likelihood of a variant sequence after marginalizing the latent variable that corresponds to the mixture assignment; the dashed blue and red lines represent the relative share of the marginal likelihood from the benign and pathogenic clusters respectively (*i.e.* the product of the marginal likelihood by each cluster).



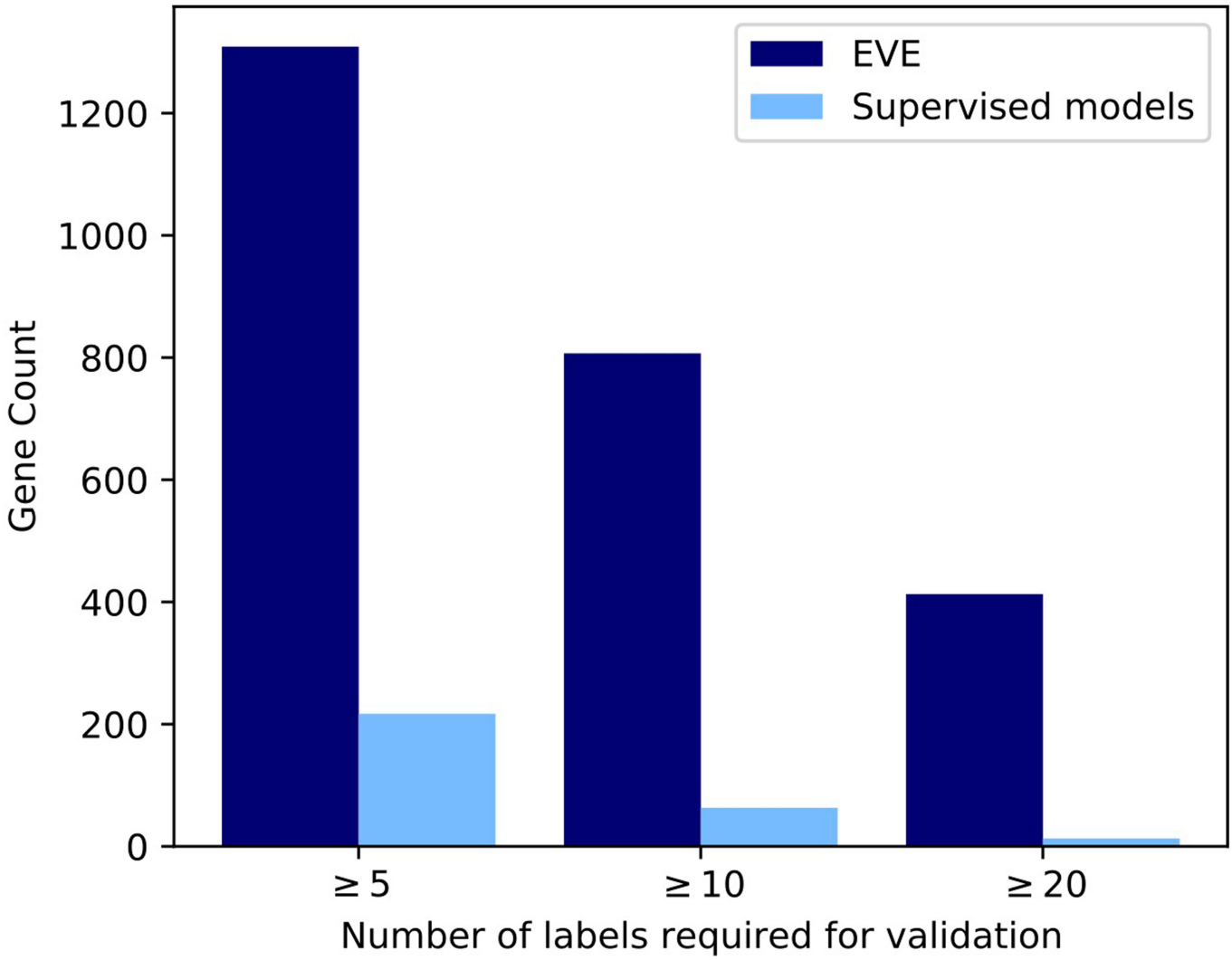
Extended Data Fig. 4 | EVE prediction for actionable genes and EVE comparison to other computational methods, including meta-predictors. **a**, EVE AUCs versus ClinVar labels for set of ACMG "actionable genes"[33] that have 15 or more labels (shown in parentheses). AUCs are computed both for EVE scores of all variants (pale blue), and of the 75% variants with most confident scores (dark blue) (Supplementary Methods). **b**, Performance comparison of EVE to state-of-the-art computational variant effect predictors: 7 unsupervised, 8 supervised, and 8 supervised meta-prediction methods. Size of marker indicates how many genes for which the method would be relevant (on a per-protein basis validation) (Supplementary Methods, Supplementary Notes 2, Fig. 2, Supplementary Tables 3, 4).



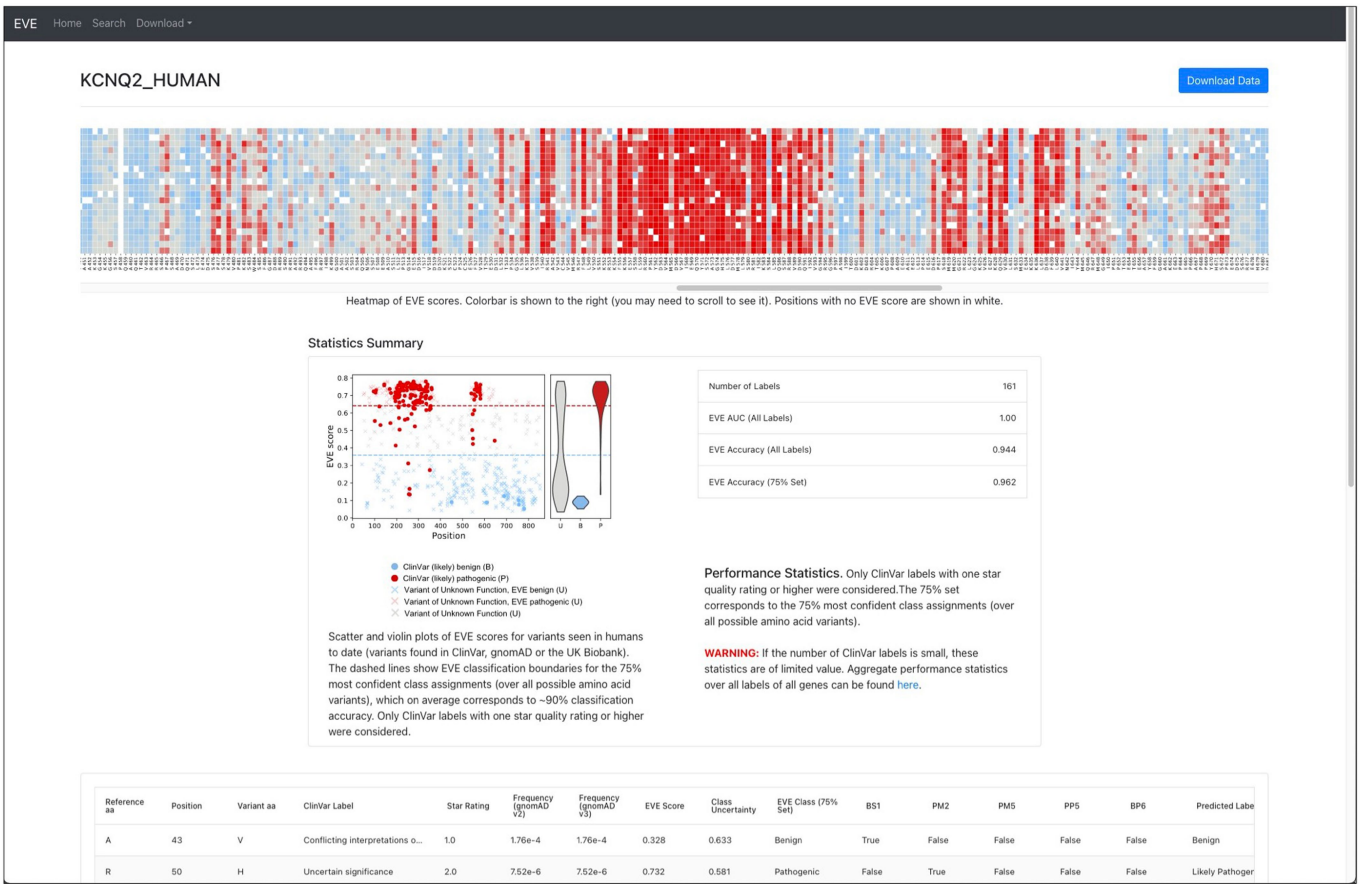
Extended Data Fig. 5 | Computational model EVE as good as high-throughput experiments for clinical labels. (Companion to Fig. 3) **a**, Comparison of computational model predictions (upper panels EVE score) and experimental assay predictions (lower panels, experimental assay metric) to ClinVar labels (dots) and VUS (crosses) and where pale red and pale blue crosses indicate EVE assignments of VUS. Dashed red and blue lines correspond to EVE predictions after removing the 25% most uncertain variants (computed on all variants across all proteins; see Supplementary Methods). x-axes are position in protein. Experimental measurements data from deep mutational scans of P53[14], from left (WT_Nutlin-3, A549_p53NULL_Nutlin-3, A549_p53NULL_Etoposide), SCN5A[13], and BRCA1[12]. **b**, Scatter plots of experiment scores (y-axis) against EVE scores (x-axis). Experimental measurements data from deep mutational scans same as **a** (Supplementary Methods, Supplementary Table 6).



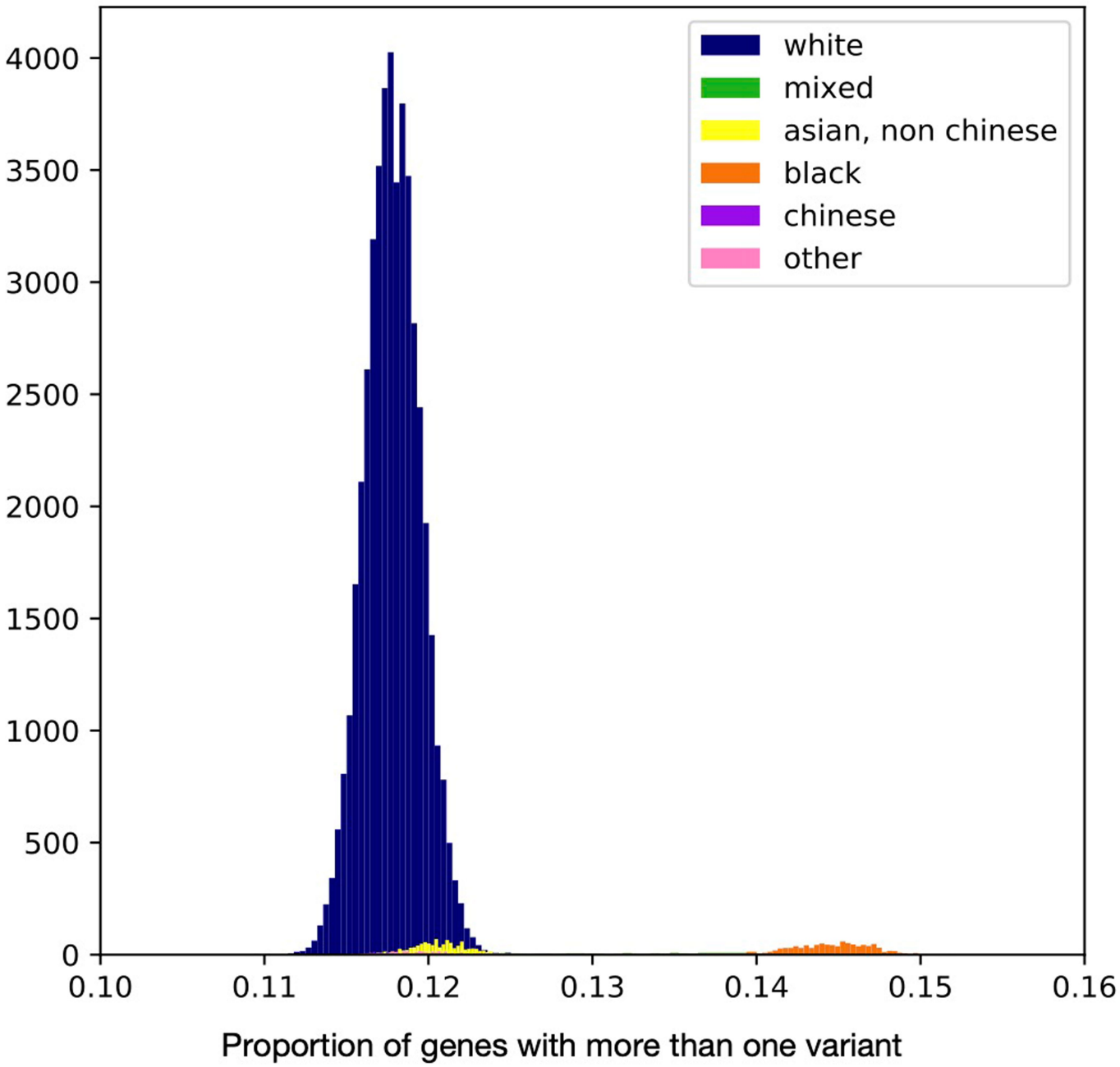
Extended Data Fig. 6 | Comparison of label policies, and comparison of EVE and experimental predictions of clinical labels. a, The y-axis is the subset of the ACMG actionable protein list with at least 5 benign and 5 pathogenic labels with at least a one-star review status in ClinVar, mean for the 3,219 proteins and mean for this subset. x-axis is AUCs computed using these labels (deep blue), labels with at least a two-star review status (light grey) and a more lenient labelling policy (sky blue), as defined in Supplementary Methods. **b**, AUC of EVE predictions (blue circle) and experimental predictions (blue cross) computed on ClinVar labels. Whilst most of the papers that provide these experimental results refer to the goal of predicting association to human disease, the assays vary in their relevance to disease phenotype. Results use high-quality labels whenever they are sufficient for robust validation (MSH2, P53, BRCA1) and lenient labels for all other cases, and 2017-release ClinVar data whenever experimental results were used in defining labels reported in 2021 (P53 and BRCA1). Reported averages of all displayed AUC values, and of AUCs computed exclusively on 2017 and 2021-reslease high-quality labels (Supplementary Methods, Supplementary Table 5,6).



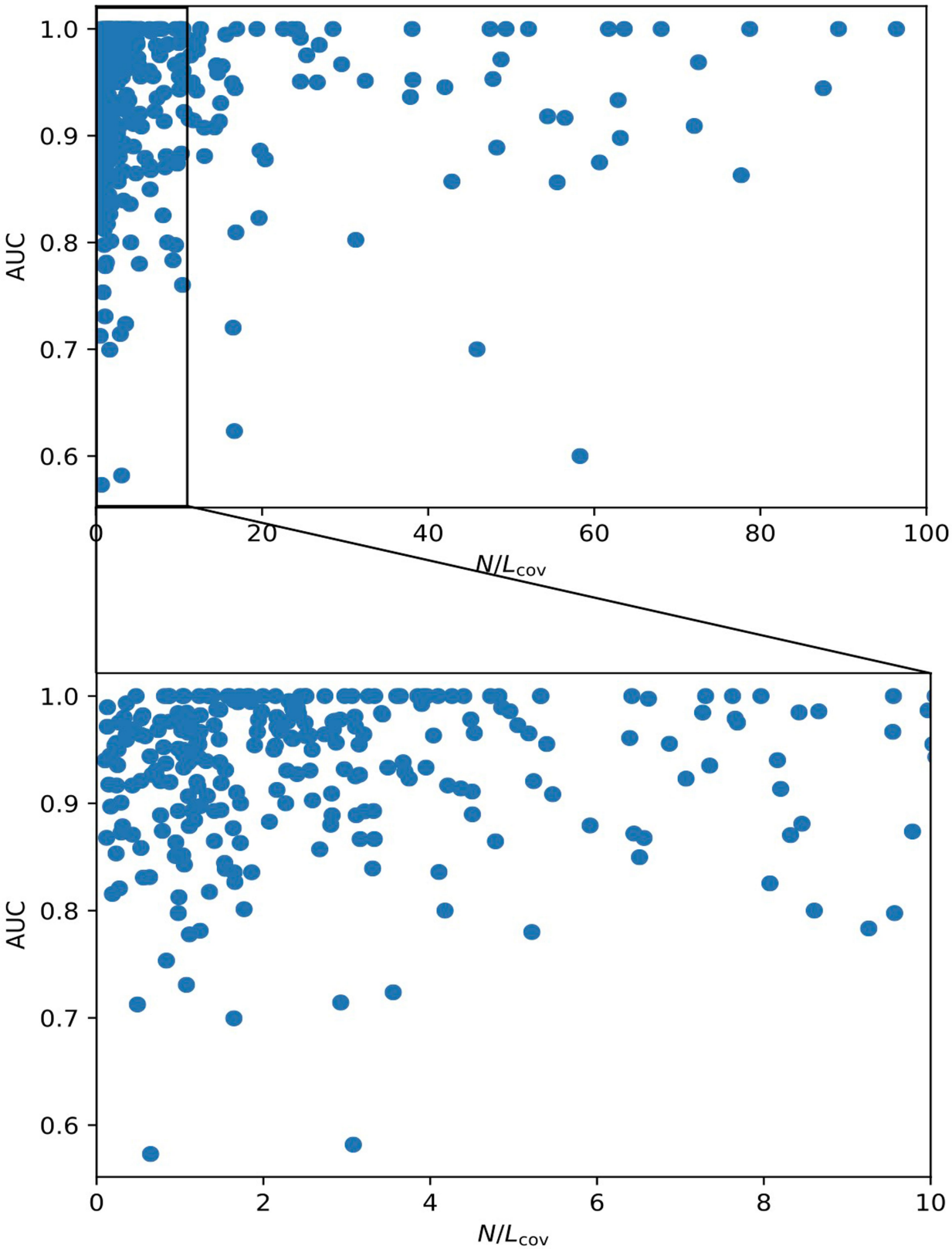
Extended Data Fig. 7 | EVE has many more genes that can be validated on, compared to supervised methods. Mean number of genes, for EVE (dark blue) and a supervised method (light blue), that have sufficient labels for validation (5 (left), 10 (middle) and 20 labels (right)). We assume a 90% train 10% test random split of all labels in ClinVar for the supervised methods.



Extended Data Fig. 8 | Data provided on our server evemodel.org. Screenshot of evemodel.org for the example of KCNQ2. Our server provides information about each protein: aggregate AUC/Accuracy, performance curves (ROC & Precision-recall), variant-level EVE scores, classification and uncertainties, as well as the multiple sequence alignments used for training. All data is available to download both in bulk and for individual genes.



Extended Data Fig. 9 | Fraction of genes per person with more than one variant. Density function of the fraction of total genes per person with at least two variants, though not necessarily in the same chromosome. Data extracted from 50k genomes of the UK Biobank with self-reported ethnicity backgrounds (Supplementary Methods).



Extended Data Fig. 10 | Performance as a function of alignment depth. Average AUC of EVE scores as a function of N/L_{cov} for the subset of genes with at least 10 known clinical labels (5 benign and 5 pathogenic). For this subset of genes, the performance of the model can be carefully validated using AUCs. There is no strong correlation between alignment depth and performance: while models with very deep alignments tend to have good performance, models with very low N/L_{cov} can also have AUC close to 1.

nature research

Corresponding author(s): Debora Marks and Yarin Gal

Last updated by author(s): Jan 8, 2021

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Multiple sequence alignments were obtained using HMMER (3.1.b2).
Data analysis	Models and analysis were written in Python (3.7) and Pytorch (1.4). All code is available through GitHub (https://github.com/OATML/EVE)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The main data analysed and generated in this study is available in Supplementary Information and at www.evemodel.org. All other data is available from original references or public repositories described in the text.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	All experimental data (and associated sample sizes) were used as published in their respective works. The 1088 genes studied in this work were chosen based on whether or not there were variants in that gene known to cause disease, and whether or not we could obtain suitable multiple sequence alignments.
Data exclusions	No data was excluded from this analysis.
Replication	All primary data used in this analysis was obtained from public repositories.
Randomization	All experimental data comes from published sources. Randomization is not relevant to the computational analysis of this study, since it is fully unsupervised.
Blinding	The same model building process was applied to all genes in this study and required no human interpretation and so no blinding was necessary.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging