*Nucleic Acids Research*, 2024, **52**, D1143–D1154
https://doi.org/10.1093/nar/gkad989
Advance access publication date: 5 January 2024
**Database issue**

# CADD v1.7: using protein language models, regulatory CNNs and other nucleotide-level scores to improve genome-wide variant predictions

**Max Schubach [1,†], Thorben Maass [2,†], Lusiné Nazaretyan [1,†], Sebastian Röner [1] and Martin Kircher [1,2,*]**

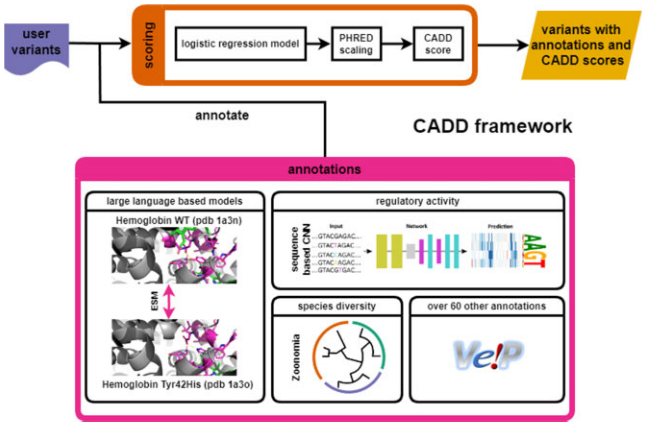
[1] Exploratory Diagnostic Sciences, Berlin Institute of Health at Charité – Universitätsmedizin Berlin, Berlin, Germany
[2] Institute of Human Genetics, University Hospital Schleswig-Holstein, University of Lübeck, Lübeck, Germany

[*] To whom correspondence should be addressed. Tel: +49 451 3101 8880; Email: martin.kircher@uni-luebeck.de
[†] The authors wish it to be known that, in their opinion, the first three authors should be regarded as Joint First Authors.

## Abstract

Machine Learning-based scoring and classification of genetic variants aids the assessment of clinical findings and is employed to prioritize variants in diverse genetic studies and analyses. Combined Annotation-Dependent Depletion (CADD) is one of the first methods for the genome-wide prioritization of variants across different molecular functions and has been continuously developed and improved since its original publication. Here, we present our most recent release, CADD v1.7. We explored and integrated new annotation features, among them state-of-the-art protein language model scores (Meta ESM-1v), regulatory variant effect predictions (from sequence-based convolutional neural networks) and sequence conservation scores (Zoonomia). We evaluated the new version on data sets derived from ClinVar, ExAC/gnomAD and 1000 Genomes variants. For coding effects, we tested CADD on 31 Deep Mutational Scanning (DMS) data sets from ProteinGym and, for regulatory effect prediction, we used saturation mutagenesis reporter assay data of promoter and enhancer sequences. The inclusion of new features further improved the overall performance of CADD. As with previous releases, all data sets, genome-wide CADD v1.7 scores, scripts for on-site scoring and an easy-to-use webserver are readily provided via https://cadd.bihealth.org/ or https://cadd.gs.washington.edu/ to the community.

## Graphical abstract



## Introduction

In recent years, the field of interpreting genetic variants has witnessed remarkable progress, driven by advancements in genomics, bioinformatics, and data analysis. The advent of high-throughput sequencing technologies has enabled researchers to generate vast amounts of genomic data, leading to an enhanced understanding of the role that genetic variants play in health and disease (1–3).

Despite significant progress, several challenges persist in accurately deciphering the complexities of the human genome. The vast number of genetic variants that are being identified makes it difficult to distinguish between neutral or benign variants and those that are clinically relevant or disease-causing. Additionally, many genetic variants exhibit subtle effects on gene function or disease susceptibility, requiring advanced computational methods and functional assays to uncover their potential impact. The genetic landscape is variable across diverse individuals and populations (i.e. haplotypes and frequencies), posing challenges in extrapolating findings about individual variants in certain patient groups (4–6).

**Received:** September 14, 2023. **Revised:** October 14, 2023. **Editorial Decision:** October 16, 2023. **Accepted:** October 17, 2023

Furthermore, our understanding of the non-coding regions of the genome has evolved, revealing their crucial regulatory roles, but also adding complexity to variant interpretation (1,7–13). The lack of comprehensive functional readouts for many variants further complicates their interpretation (3,11,14–17). Finally, ethical and privacy concerns surrounding genetic data sharing and analysis must be carefully navigated (18,19).

Combined Annotation Dependent Depletion (CADD) is a Machine Learning-based scoring system used to predict the deleteriousness or functional impact of genetic variants in the human genome (20). It integrates various genomic annotations and functional information to assign a single numerical deleteriousness score to each variant, correlating with the likelihood that the variant is pathogenic or disruptive to gene function. CADD is based on logistic regression models and applicable to single nucleotide variants (SNVs) and short inserts and deletions (InDels), throughout the human genome reference assembly.

CADD makes use of a wide range of features, including DNA sequence properties, gene and transcript models, scoring of protein-coding effects, conservation across species, biochemical activity and other genomic annotations. By combining multiple sources of information, CADD aims to provide a more accurate prediction of variant deleteriousness compared to using individual features. Further, by integrating the partially correlated features in one score, it also overcomes issues from individually considering multiple annotations and presuming them as independent evidence for a variant effect. CADD has been widely adopted in the field of genomics and variant interpretation, aiding researchers and clinicians in prioritizing genetic variants for further analysis, especially in the context of disease-causing variants and personalized medicine applications (21–23). The principle behind CADD has been used to develop scores for further species (24–26), to infer measures of selection (27) and constraint (28) and extended to further variant classes, specifically structural variants (29). Further, the broader concept of using species differences and long-standing variation has been applied to develop improved models of missense effects (30,31).

As our understanding of genetics and genomics is rapidly evolving, CADD aims at continuous integration of new research discoveries to provide the most accurate predictions and to support the latest human genome builds. Since the CADD v1.6 release in 2021 (32), which specifically focused on improving the scoring of splice effects using sequence-based models, new methods for the identification of functional regions and the assessment of impact of variation in the human genome have been developed. In CADD version 1.7, we again added new features that improve CADD scores for certain variant effects (Figure 1), boosting the overall performance of CADD and bringing new developments to the community.

Specifically, deep learning methods that frequently outperform other models have been of continued interest to the community. Here, we include scores derived from the Evolutionary Scale Modeling (ESM) for assessment of variants in protein coding regions (33) as well as from a convolutional neural network (CNN) trained on open chromatin sequences, as a proxy for regulatory regions in the genome. Further, we complement the previously included conservation scores with the updates of the Zoonomia project (34) and include new annotations for 3′ Untranslated Regions (3′ UTRs) (35) as well as models of genome-wide mutational rate (36). Finally, we update our gene and transcript models by advancing from Ensembl version 95 to 110 and using an updated version of Ensembl Variant Effect Predictor (VEP). As for previous CADD versions, genome-wide scores, scripts for on-site scoring and an easy-to-use webserver are readily provided to the community for CADD v1.7 (Figure 1).

## Unique characteristics of the CADD framework

CADD is different from many other machine learning methods for variant prioritization that are frequently trained on the small number of genetic variants that have a known effect on health or disease (37–40). Instead, CADD uses a larger and less biased set of training data. It assumes that most of the variants that have appeared and stayed in humans after speciation are harmless or neutral, because they have survived millions of years of natural selection; we call these variants ‘proxy-neutral’. These variants are compared with a set of simulated variants that have not been purified by selection; while many of these variants are neutral, some of them would be harmful or have some effect if they occurred in a person; we call these variants ‘proxy-deleterious’. The main idea of CADD is to contrast the proxy-neutral and proxy-deleterious sets based on available genomic annotations using a machine learning approach, i.e. to learn that there are fewer harmful mutations in the proxy-neutral set by respective shifts in the annotation features. Thus, creating a model for estimating the deleteriousness of all types of short sequence variants, coding and non-coding, across the human genome.

The proxy-neutral set consists of more than 14 million human lineage derived sequence alterations (14 million SNVs and 1.7 million InDels up to 50 bp), inferred from primate whole genome alignments (41). The proxy-deleterious variation is a size matched set of simulated variants. We simulate these ‘*de novo*’ variants according to the substitution frequencies and insertion/deletion lengths observed in the proxy-neutral set, accommodating a local adjustment of mutation rates and asymmetric CpG-specific mutation rates (20). The resulting set of about 30 million variants is annotated with more than 100 annotations derived from gene model information, sequence conservation and constraint, epigenetic and regulatory activity, variant density, and others.

In all recent CADD implementations, we train a logistic regression model on this data. To account for some non-linear relation between annotations, CADD creates crossed feature annotations (20), like including conservation scores or epigenetic activity model coefficients based on the annotated consequence labels (e.g. missense, nonsense, upstream, downstream, intergenic). The resulting model is then applied genome-wide, i.e. we compute the logistic regression scores (raw scores) for all possible single base pair alterations of the human genome. Based on these genome-wide raw scores, a conversion table is generated that translates to a scaled CADD-score, a variants’ relative rank among all potential SNV alterations presented on PHRED-scale ($-10 * \log_{10}(1/\text{rank})$). The conversion table is used to obtain scaled scores for any SNV, multi-nucleotide substitution or InDel change, representing the commonly used CADD scores (abbreviated as C-scores).

As raw scores are the immediate output from the machine learning model, they summarize the extent to which the variant is likely to have derived from the proxy-neutral (negative values) or proxy-deleterious (positive values) class. Raw scores have only relative meaning with higher values

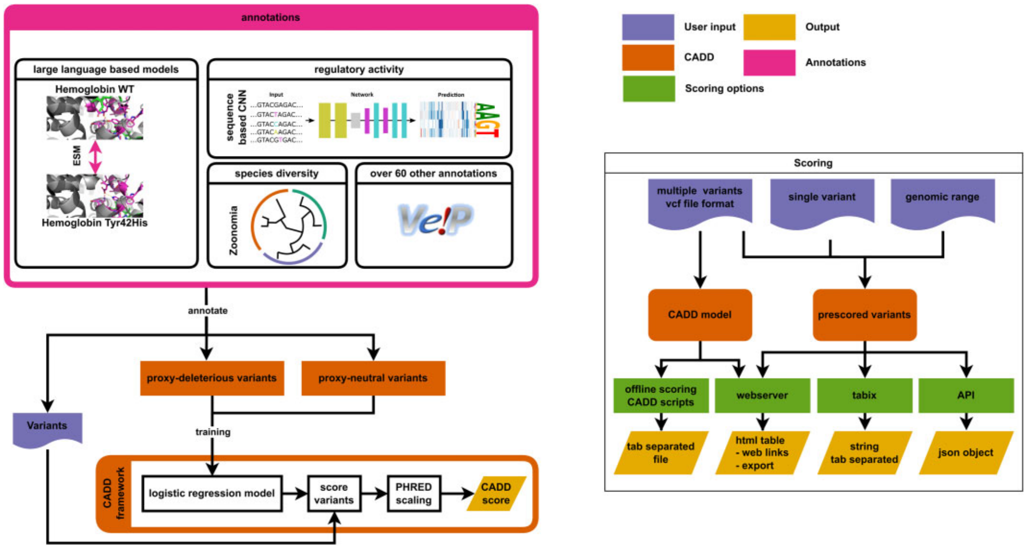


**Figure 1.** Combined Annotation Dependent Depletion (CADD) workflow. CADD is a Machine Learning-based scoring system for predicting the deleteriousness of genetic variants across the genome. It integrates various genomic annotations and functional information in a logistic regression model to assign numerical deleteriousness scores for all possible single nucleotide variants (SNVs), multi-nucleotide substitutions and insertion/deletion changes (InDels). From a user's perspective, a provided variant list (typically in Variant Call Format, VCF) is annotated with diverse annotation features. Combined CADD scores are determined and returned for the provided variants with or without the additional annotations. In CADD v1.7, we integrated new annotation features, among them state-of-the-art protein language model scores (Meta ESM-1v), regulatory variant effect predictions (from sequence-based convolutional neural networks) and sequence conservation scores (Zoonomia). Various sources are available for obtaining CADD scores. CADD models and offline-scoring are available to calculate and retrieve scores on-premises. A webserver offers the download of pre-scored SNV and InDel variants, a VCF upload for online-scoring and exploring pre-scored SNV variants directly on our website. Pre-scored variants are also accessible through HTSlib/tabix and API calls. PDB entries 1a3o and 1a3n were used to create graphical representations of Hemoglobin shown in the left-most panel.

indicating that a variant is more likely to have derived from the proxy-deleterious than the proxy-neutral variant set. However, they change between distinct annotation combinations, training sets or tuning parameter choices (i.e. CADD model versions) and thus do not have absolute meaning and cannot be compared across models. Raw scores offer superior numerical resolution and preserve relative differences between scores that may otherwise be rounded away in the scaled *C*-scores. Thus, when statistically comparing score distributions between groups of variants (e.g. in cases versus controls), raw scores should be used (42).

In contrast, 'PHRED-scaled' C-scores are normalized to all potential ∼9 billion SNVs, and thereby can be compared across model versions and reference builds. Regardless of the annotation set or model parameters, a scaled score of 20 or greater indicates a raw score in the top 1% of all possible reference genome SNVs, and a score of 30 or greater indicates a raw score in the top 0.1%. When identifying causal variants or performing a fine-mapping of variants within associated loci, scaled scores are advantageous as they allow the user a direct interpretation in terms of the estimated deleteriousness relative to all possible SNVs in the reference genome.

Since its inception, we have been advocating for ranking variants by CADD scores rather than declaring a single universal cut-off value to declare a variant 'pathogenic' (or 'functional' or 'deleterious') as opposed to 'benign' (or 'nonfunctional' or 'neutral'). We believe that such binarization is flawed due to its substantial loss of information and as the choice of the cut-off would naturally depend on specific factors, such as the severity of the phenotype and how much expression variation is tolerated (e.g. haploinsufficiency, dominance, recessiveness), or the amount of time and resources available for curation or experimental follow-up (42).

## Updates to the CADD annotation set

The CADD framework is highly flexible and modularized, enabling further improvements, like the integration of new annotations as the result of scientific developments. Despite its imperfect approximation of training set labels, the key advantage of the CADD framework is the comprehensive and systematic labeling of tens of millions of variants for the training set. Each iteration of the CADD model is therefore trained on about 30 million variants and hundreds of features derived from available annotations. This enables CADD to accommodate nearly any feature that can be tied to reference assembly coordinates, and the capacity to score both coding and noncoding variants. The size of the training set allows integration of many annotations without substantial risk of overfitting.

To explore potential updates to CADDs feature set, we reviewed recent literature for variant scores and annotations that could potentially improve the performance of CADD (Supplementary Table S1). Among these scores and annotations are APARENT2 (35), Zoonomia (34), Roulette (36) and gwRVIS scores (43). APARENT2 is the successor of APARENT, a sequence-based deep learning tool that quantifies a variants' potential to disrupt alternative polyadenylation (44). The Zoonomia project provides conservation information for each position in the genome by comparing the genomes of more than 200 mammalian species, substantially improving on CADD's previously used 43 mammalian sequences (34). The increased number of species comes with higher resolution on the level of sequence conservation scores and more complete coverage along the genome. In previous versions, conservation has been a major indicator of which regions and nucleotides are prone to disease causing effects. It is also one of the most important predictors in the sparsely annotated non-coding genomic regions (45). Roulette is a sequence mutability score, i.e. a score reflecting the sequence and its context dependent propensity to acquire *de novo* mutations (36). GwRVIS is a genome-wide score that quantifies intolerance to variation with nucleotide resolution. Conservation information is however explicitly excluded, distinguishing it from phylogenetic scores (43).

We also derived scores from ESM-1v protein language models (33) for single amino acid substitutions, inframe insertions and deletions, as well as frameshifts and stop gains (see Supplementary Note S2 for more details). ESM-1v is a transformer protein language model developed by the Meta Fundamental AI Research Team that aims at predicting effects of missense variants. The 650 million parameter model has been trained directly on protein sequence databases in an unsupervised manner. It has been proposed that the model understands protein function on a molecular-to-atomic level and was shown by several groups to be one of the best performing tools for missense variant prediction (33,46,47).

In the field of predicting epigenetic properties of non-coding elements, significant progress was made by the development of deep learning architectures. Here, Enformer is the most prominent example as well as the most complex architecture with a wide sequence context around each central position (48). It was shown that Enformer can effectively predict epigenetic marks from DNA sequence alone and allows to infer the sequence motifs underlying functional sequence alterations. From these and other modeling efforts, open-chromatin information like DNase-seq or ATAC-seq seem to provide good approximations for predicting regulatory variant effects, for example the activity measured from massively parallel reporter assays (MPRAs) (48–51). However, due to Enformer's complexity, its many cell-type specific outputs, and its computational time requirements, it is not feasible for us to apply it genome-wide and to integrate it into CADD. Instead, we decided to train a small-scale convolutional neural network (CNN) model on open chromatin sequences from multiple cell-types (see Supplementary Note S3 for details). Depending on the benchmark MPRA data set, our model showed a performance comparable to Enformer with much faster calculation times, motivating us to test its integration into CADD.

To evaluate whether the new annotations improve CADD's performance, we collected several suitable benchmark data sets (see Supplementary Note S1). Our previous CADD model (CADD v1.6) served to determine a baseline performance on the benchmark sets and was compared with CADD models with one of the new annotations added (potentially resulting in multiple new CADD model features being incorporated). Further, we included the respective annotations or features as a standalone score in our benchmarks. We considered a significant improvement of the individual annotation added to CADD over the baseline model as sufficient criterion for integrating the respective score or annotation in our new CADD v1.7 model, which then included all new features.

## Protein language models

The ESM-v1 language model for protein coding sequences consists of five models (initiated with different seeds during training) whose outputs are typically averaged. Provided with only an amino acid (AA) sequence, the models return likelihoods for specific AAs at each position along the sequence. To score missense variants, we used the average log odds ratio between the alternative and reference amino acid across the ESM-v1 models in a window of 350 AAs (see Supplementary Note S2 for details). To test the impact of the ESM-1v-derived missense score, we used Deep Mutational Scanning (DMS) data sets of human proteins from the ProteinGym database (52). We considered all amino acid substitutions that can occur due to a change of a single base pair and obtained about 41 000 DMS scores corresponding to about 31 000 variants (see Supplementary Notes S1 and S6). Next, we correlated DMS scores with CADD scores (Figure 2A). By including the ESM-1v-derived missense score in our CADD model, we observe an improvement for 29 of the 31 data sets and an overall improvement from the Spearman correlation of 0.217 ± 0.006 for the CADD v1.6 baseline model, to 0.250 +/- 0.006 for the model with ESM (Figure 2A). Of note, experimental DMS scores and our ESM-derived scores provide information on the effect of a variant on the protein level only, excluding for example splice effects. The extended CADD model does not reach the performance of the ESM-1v-derived missense score on these DMS effects. However, testing the scores on clinically relevant variants from ClinVar with two and more star reviewer status (abbreviated as **+; see Supplementary Note S1 and S6) shows that CADD outperforms ESM-1v-derived missense scores (Supplementary Figure S1a), when multiple molecular effects integrated in CADD can contribute to a variants' deleteriousness.

For the ESM-1v-derived score for inframe insertions and deletions (modeled as the log odds ratio of reference vs. alternative sequence in a context of 250 AAs, see Supplementary Note S2 for details), we used variants from two previously collected test data sets (53). In brief, the first data set contains benign and pathogenic variants from gnomAD v3.1 (54), ClinVar (37), and the Deciphering Developmental Disorder (DDD) study (55), resulting in a total of 2088 benign and 1655 pathogenic variants (see also Supplementary Notes S1 and S6). We observe an increase in area under the receiver operating characteristic curve (AUROC) values from 0.85 to 0.88 for our CADD model containing the ESM-1v-derived inframe insertion and deletion scores (Figure 2b). Here, the ESM-1v-derived score as a stand-alone predictor does not reach the performance of our CADD models. The second data set is a subset containing only variants from the DDD study (80 benign and 70 pathogenic variants; see also Supplementary Notes S1 and S6). We also report a positive trend for this data set, but note that the uncertainty associated with AUROC values is relatively large due to the small number of variants (Supplementary Figure S1b).

We also derived an ESM-1v score for frameshifts and stop gains by considering the median effects of lost amino acids in the log ratio (see Supplementary Note S2 for details). However, validation of the score proved difficult as we were not able to identify a suitable set of benign frameshift and stop gain variants to assess the performance. Pathogenic variants (**+ ClinVar reviewer status) were readily obtained using ClinVar InDels resulting in frameshifts (11 574 variants, see Supplementary Note S1 and S6) or SNVs resulting in stop gains (8159 variants, see Supplementary Note S1 and S6). In generating a benign variant set from ClinVar, we identified only 3 stop gains from InDels <50 bp, 15 stop gains from SNVs, as well as 44 frameshifts (see Supplementary Note S1). Testing our score on these data sets showed a positive trend but no significant improvement of the corresponding AUROC values when including the ESM-1v-derived scores

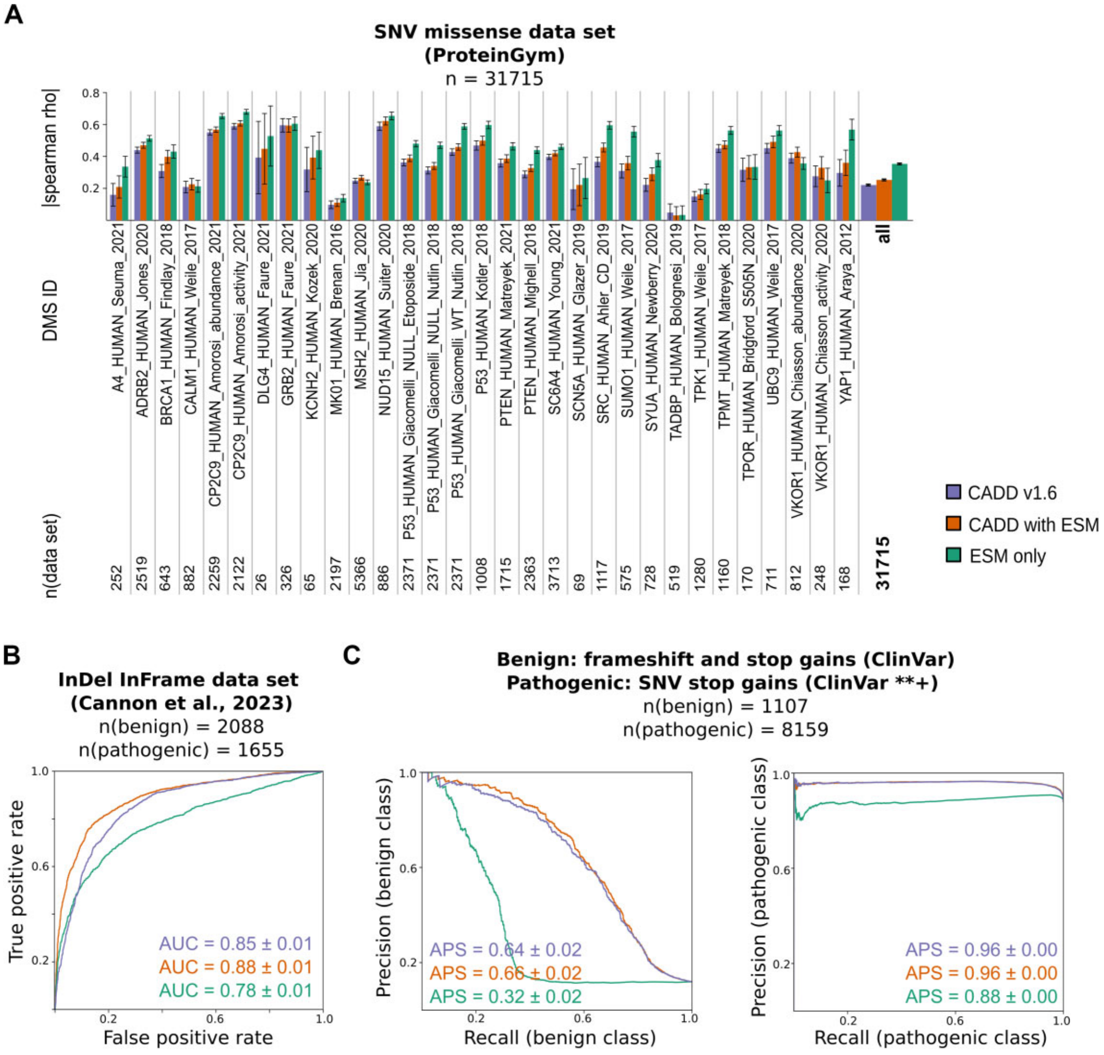


**Figure 2.** Including ESM protein language model scores improves the performance of CADD in coding regions. (**A**) Spearman correlations of CADD v1.6, a CADD model including ESM protein language model scores for missense variants, and the stand-alone ESM protein language model score for missense variants with experimental effect scores form the ProteinGym (52). (**B,C**) Receiver operating characteristics (ROC) curves and precision recall curves with the corresponding area under the ROC (AUC) and average precision score (APS) values. These metrics are shown for CADD v1.6, a CADD model including our ESM score for inframe InDel variants (B), for stop gain or frameshift variants (C), and for the derived ESM scores as stand-alone predictors (B, C). See also Supplementary Figure S2a for ROC curves of the data set shown in (C). The positive class corresponds to the pathogenic class if not stated otherwise.

(Supplementary Figure S2). Using ClinVar stop gains and frameshifts, independent of their review status for the benign set, gave us a total number of 1107 variants (see Supplementary Notes S1 and S6), which together with the pathogenic ClinVar **+ variants resulted in a more balanced test data set and confirmed the positive trend (Figure 2C). Of note, we also tried to utilize gnomAD v3.1 to obtain benign variants but filtering for frameshift and stop gain variants with an allele frequency above 0.5% resulted in less than 400 variants. In summary, the applied validation sets did not result in a measurable performance gain for frameshifts and stop gains, but we still included the ESM-derived scores for these variant classes in CADD v1.7 based on the positive trend.

## Regulatory sequence models

We trained a CNN model that predicts probabilities for variants being part of a regulatory element. The model was trained on DNase-seq data of seven different well studied ENCODE cell lines (see Supplementary Note S3, Figures S3 and S4, Table S2 for more information) and GC-matched background sequences as negative sample to improve the multitask performance of the model (56). The structure and parameters of the deep neural net were optimized, resulting in the final model (called RegSeq) which uses 500 bp of input sequence centered at the variant and has three convolutional layers and two dense layers (see Supplementary Table S3 for hyperparameter optimization and Supplementary Figure S5 for a schematic model overview).

We benchmarked our model with experimental reporter assay regulatory variant activity readouts from saturation mutagenesis MPRAs of different promoter and enhancer elements (57) (see Supplementary Note S3, Supplementary Figures S6 and S7). RegSeq showed an improved prediction over CADD v1.6 using cell-type specific, agnostic (average over 7 cell lines), and GC-matched background variant effect predictions. Pearson and Spearman correlations were similar (Supplementary Figures S6 and S7) to the performance of the larger Enformer model (48), which is based on a transformer architecture and more than 100 kb sequence context. Enformer is currently one of the best performing tools on variant effect prediction for regulatory sequence effects (48,58,59) and it would have been our preferred target for integration in CADD. However, the slow prediction time per variant effect (around 4 s per variant on one GPU) made it unreasonable to score the CADD training data and nearly impossible to create genome-wide predictions, motivating our decision to build and integrate the considerably smaller model (RegSeq).

We integrated the cell-type agnostic (average) RegSeq variant effect prediction as well as the GC-matched background predictions in CADD using several transformations on positive (activating) and negative (repressive) output predictions (e.g. minimum and maximum over cell-types; see Supplementary Note S3 and Supplementary Table S4). We trained a new CADD model with the additional RegSeq features and correlated the output with the significant experimental effects from the saturation mutagenesis data (Supplementary Note S1). Figure 3A shows the Spearman correlation of all variants (1 bp deletions and SNVs, $n = 4332$) per MPRA element before and after integration of RegSeq features (see correlation of all variants across elements in Supplementary Figure S8 and S9). We see an improved performance over nearly all elements as well as an overall improvement over the regulatory variant effects. Figure 3B shows a subpart of the promoter of the gene *Factor IX (F9)* with a binding site of *ETS*-related factors. The RegSeq model as well as the extended CADD v1.6 model with its integration can recover the binding site. We conclude that new sequence-based regulatory features are effectively used in the logistic regression model and result in improved regulatory variant effect predictions.

## 3′UTR, non-coding constraint, mutational scores and conservation scores

We and others have observed that CADD scores have good performance in coding regions of the genome whereas deleterious non-coding variants are more difficult to analyze with it (60–63). We believe that this is mostly due to a limited number of predictors of molecular effects outside of coding sequence (13,17,32). To improve CADD's performance, new annotations targeted at non-coding variants were explored. First, we considered a sequence-based deep-learning score, APARENT2, quantifying the impact of genetic variation on transcript polyadenylation and 3′ cleavage (35). To assess the performance of CADD models with and without APARENT2, we tested them on pathogenic and benign 3′UTR variants from ClinVar. We only used variants for which a precalculated APARENT2 score was available (i.e. 23 pathogenic and 3865 benign variants, see also Supplementary Notes S1 and S6). We observe a significant improvement of average precision scores (APS) from 0.105 to 0.381 (Figure 4A). However, due to the small set of pathogenic variants, the uncertainty is rather large (standard deviation of 0.06).

Next, the genome-wide residual variation intolerance score (gwRVIS) was tested as an additional non-coding feature (43). It models intolerance to variation from functional genomic annotations and primary genomic sequence to prioritize noncoding variants. Notably, the score does not use conservation information, which, its authors argue, should better reflect human-specific constraint regions. To assess the impact of integrating gwRVIS in CADD, we used 3′UTR and 5′UTR variants as well as non-coding transcripts and intergenic variants from ClinVar for which a gwRVIS score was available (see also Supplementary Notes S1 and S6). We observe little predictive power of gwRVIS on this evaluation data set and a negative impact on CADD when adding this score into the feature set (Supplementary Figure S11), which is why we did not include gwRVIS in the new CADD release.

We further investigated the impact of the genome-wide mutation score Roulette (36). It models local mutation rate along the genome by including known determinants like the extended sequence context of 6 upstream and 6 downstream nucleotides, methylation, expression, transcription and replication directions and the observed mutability of each trinucleotide context in 50 kb windows. To assess the impact of Roulette, we needed a genome-wide measure of variant constraint and used variant minor allele frequency as a proxy. We therefore calculate the Pearson correlation coefficients of the respective CADD models with minor allele frequencies derived for a random subset of SNV and InDel variants ($n = 100\,000$ each) in the 1000 Genomes project (64). We observe an increase in the absolute value of the correlation coefficients, especially for SNVs from 0.045 to 0.058 when including Roulette (Figure 4b).

As outlined above, we also explored the impact of the higher depth Zoonomia alignments and included metrics derived from the 43 primate genome and 241 mammalian genome alignments (34). This included phyloP and phastCons conservation scores available at base-level resolution, a set of annotated Ultra Conserved Elements (UCE, at least 235 species aligned and all aligning species are fixed for the same base at every position) and a set of Runs of Contiguous Constraint (RoCC, genomic regions where contiguous bases have phyloP score >2.27 in the mammalian alignment). When tested on correlation with minor allele frequencies as described for Roulette above, we observe an increase for SNVs from 0.045 to 0.066 when including Zoonomia conservation scores (Figure 4b).

## An updated CADD model

After validating the impact of the above-presented individual features, we combined them and added them to the features already included in CADD v1.6. We further updated the Ensembl transcript database as well as the Ensembl VEP tool (65,66) from version 95 to the most recent version 110 and trained a new CADD model (CADD v1.7). The model was tested and compared to CADD v1.6 on all ClinVar variants annotated with either benign ($n = 35\,037$) or pathogenic ($n = 26\,981$) clinical assertion with a two star and better reviewer status. In an additional assessment, we substituted the benign variants with variants from ExAC with a minor allele frequency ≥5% ($n = 69\,543$; see Supplementary Note S1 and S6). On this set of variants, predominantly in coding genes, we see slight but significant improvements in AUROC values from 0.981 to 0.982 or from 0.985 to 0.986 (Supplementary Figure S12), respectively. Even though we are aware that variants with clinical assertions in the ClinVar database come with their issues and reflect by no means the entire genome (13,67), we think that this test unambiguously shows the improvement of CADD with the incorporation of new annotations.

CADD has been compared in many benchmarks and to many other tools in the past (32,46,47), which is why we are not presenting an extensive benchmarking here. Benchmarks

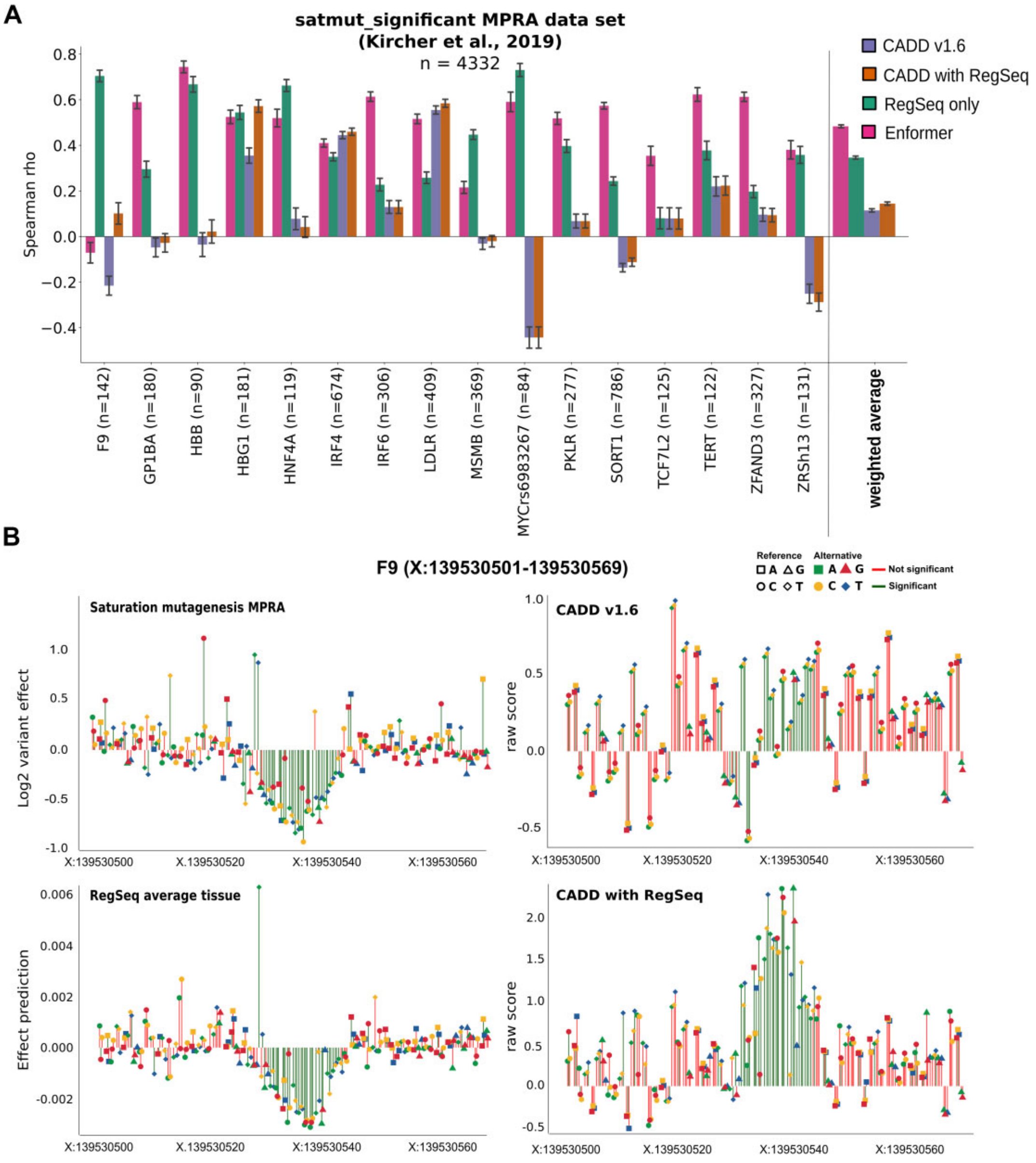


**Figure 3.** Integration of RegSeq features within CADD. Panel (**A**) shows the Spearman correlation of significant saturation mutagenesis reporter readouts (1 bp deletions and SNVs with min. 10 barcodes and *P*-value $< 10^{-5}$) per element (57) with CADD v1.6 with and without integrated RegSeq features, an individual RegSeq feature (average across cell-types) and Enformer's average DNase-seq predictions (48). The absolute value of the MPRA effect is used for CADD, as it is not expected to predict the directionality of the expression effect. Weighted average is the mean Spearman correlation weighted by the number of variants in each element. Bars show the average correlation after bootstrapping (1000 runs with 80% of element variants) and error bars represent the standard deviation. Other data sets and Pearson correlation values are available in Supplementary Figures S8 and S9. Panel (**B**) shows part of the *Factor IX (F9)* promoter, where mutations lead to a significant reduction in activity due to association with *ETS*-related factors (57). The RegSeq model (lower left) aligns well with the variant effects measured by MPRA (upper left). CADD v1.6 with integrated RegSeq features aligns better with the motif (lower right) compared to the original CADD v1.6 (upper right). The whole *F9* promoter region is shown in Supplementary Figure S10.

are frequently limited to certain molecular functions or also use variants with clinical assertions, which brings up another frequently encountered limitation, namely that some tools are explicitly trained using ClinVar or similar variant sets. This makes it impossible to assess their performance on clinically relevant variants in an unbiased way, as there are no independent data sets that would not share the historic ascertainment or other basic characteristics of the ClinVar database. It should also be noted that distinguishing methods by their means of training only (e.g. supervised vs. unsupervised learning approaches) is insufficient and that the training objective needs to be considered (46,47). The inclusion of features derived from training on pathogenic variants may not immediately render a method unsuitable for comparison but may just complicate its benchmarking, as variants used in training the included feature may need to be eventually excluded. However, CADD integrates PolyPhen missense scores as a feature, a method that was trained on (a comparably small number by current standards) known pathogenic variants, while CADD's training objective is the separation of human-derived from simulated variants. CADD's initial publication (20) therefore included a benchmark of a model with and without PolyPhen scores, showing that CADD's performance is not tainted by the inclusion of PolyPhen scores.

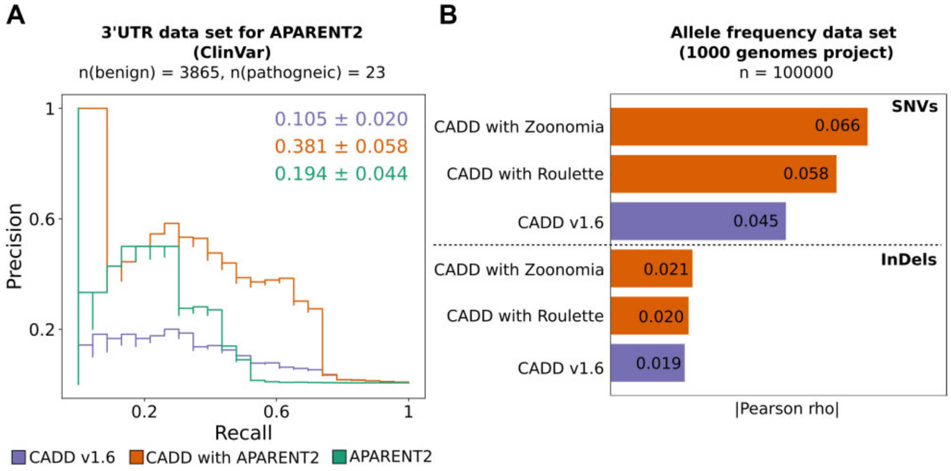


**Figure 4.** Comparison of CADD performance with and without APARENT2, Roulette, and Zoonomia features. (**A**) Precision recall curves and corresponding Average Precision Scores (APS) values of CADD v1.6, CADD including APARENT2 (35) and the absolute value of the APARENT2 score. The positive class corresponds to the pathogenic class. (**B**) Pearson correlation coefficients of CADD scores from models with and without Roulette (36) or Zoonomia (34) features with minor allele frequencies of 100 000 SNV and InDel variants sampled from the 1000 Genomes Project (64).

Another challenge of these benchmarks is that ClinVar and other clinically ascertained variant sources remain largely constrained to variants in and around protein-coding variants, despite the increased application of genome sequencing over the last few years (13). This is caused by our still limited understanding of variants in regulatory sequences and the large computational and experimental burden of analyzing the other approximately 98% of the genome (16,68). Additionally, specialized tools may perform better on specific benchmarking tasks, while scores like CADD integrate across molecular functions and provide genome-wide predictions (32). This makes fair comparisons difficult, as shown by the analysis of CADD and ESM-1v scores above. When comparing the two tools on protein function specific experimental assays (e.g. DMS data; Figure 2A), ESM-1v clearly outperforms CADD. ESM was designed to learn basic principles of protein sequence to structure and function relations, without the specific genomic encoding involving splicing, codon usage or for example mutational processes. Thus, on a set of missense variants that may have effects along genomic encoding, transcription, translation and protein function (e.g. ClinVar missense SNVs, Figure S1a), CADD outperforms ESM-1v. For such a data set, CADD profits from the many genomic annotations and features that contribute to variant deleteriousness.

## Web access and score availability

CADD and its associated software are freely available for all non-commercial applications and otherwise currently licensed through the University of Washington, Seattle. CADD scores are available for SNVs as well as InDels shorter than 50 bp located on the 22 human autosomes and chromosome X. We further provide scores for chromosome Y, although not all model annotations are available. Due to a lack of available annotations and consistent consideration across various studies, we do not support alternative haplotypes and other contigs. Due to differences in inheritance, gene density, transcription machinery and the availability of annotations, we no longer support scoring of mitochondrial variants in CADD v1.4 and later versions. While CADD will return scores for variants 50 bp and longer, this is outside of the range that it was trained on and scoring should be done with dedicated tools like CADD-SV (29).

CADD scores can be accessed in various ways, including through a number of third-party sources, such as dbNSFP (69), as a plug-in for Ensembl VEP (66), ANNOVAR (70), SeattleSeq (71), ExAC/gnomAD (54,72) and PopViz (73), but we recommend users to access information, scripts and pre-scored files primarily from our US (https://cadd.gs.washington.edu/) or Germany-based (https://cadd.bihealth.org/) webservers (see also Figure 1). These servers allow retrieval of SNV scores by position or in ranges, on the website or through a REST-API. In order to enable external sources to refer directly to CADD scores on our webservers, we implemented direct and versioned links to the scores of SNVs (e.g. https://CADD-SERVER/snv/BUILD-VERSION_inclAnno/CHROM:POS_REF_ALT). Genome-wide pre-scored SNV files as well as select pre-scored InDels are available for download. Additionally, VCF files with up to 100 000 positions can be uploaded and annotated with CADD scores, and users are informed by email once the processing is finished and variants can be downloaded. All scripts and annotations required for scoring SNVs and InDels are linked and available in the public GitHub repository (https://github.com/kircherlab/CADD-scripts).

We provide software for offline scoring especially for users that are legally required to score SNV and InDel variants on their own systems (18,74). Offline scoring takes a VCF file as input and allows for retrieval of annotations from pre-scored variant sets (to reduce computational time), as well as direct annotation and scoring of the remaining variants. It returns a gzip-compressed tab-separated text file (tsv.gz) containing all scored variants, with or without annotations. Offline scoring is based on the workflow management system Snakemake (75) with dependency management through conda (https://conda.io). We provide an installation script that downloads all necessary annotations and, optionally, pre-scored variants.

Further, we provide bigWig files of the maximum SNV score per genomic position that can be visualized as browser tracks for utilities like the UCSC genome browser or Integrative

Genomics Viewer (IGV) and allow users to screen larger genomic areas quickly. Finally, the proxy-benign and proxy-pathogenic variant sets, the annotated training data as well as a comprehensive set of test and validation sets are available for other research efforts (https://cadd.gs.washington.edu/training or https://cadd.bihealth.org/training).

## Discussion and future work

We present a new release of the Combined Annotation Dependent Depletion scores (CADD version 1.7) that integrates annotations from recent community efforts on the assessment of variant effects, as well as new conservation and mutation scores. Over recent years, various deep learning methods have shown their potential to outperform other kinds of models and we include deep learning scores derived from Evolutionary Scale Modeling in protein coding regions (33) as well as a CNN model of open chromatin regions for the effects in regulatory regions. Further, we included conservation scores with five times more species from the Zoonomia project (34), a new annotation for 3′UTRs (35), and models of genome-wide mutational rates (36). Finally, we updated our gene and transcript models and version of Ensembl VEP (65,66).

For about a decade, CADD has been one of few machine learning methods able to provide genome-wide scores for SNVs, multi-nucleotide substitutions and InDels. Even though most known variants with a clinically relevant effect are in coding regions, it is probably the non-coding part of the genome that is key to delivering diagnoses to the many undiagnosed patients as well as understanding gene expression and regulation of all molecular processes of the cell. This is also not a new realization. For example in 1975, it was described that protein-coding differences between humans and chimpanzees were insufficient to explain their phenotypic differences (76). A recent study substantiated this and found only 126 out of 24 374 human-specific variants that are coding missense variants (77). Thus, we must presume that future versions of CADD will continue to put efforts in improving the deleterious predictions of non-coding variant effects.

While we show that integrating carefully selected annotations into CADD can improve its performance to predict deleteriousness, the model complexity (and computational expenses) increases with each new annotation. Especially deep neural network derived annotations, like the large language model ESM, are optimized for GPUs and the respective infrastructure is required to perform predictions on a genome-wide scale. We have taken up much of this computational burden by pre-calculating scores for all potential SNVs, but for users it still complicates running CADD scoring on their own infrastructure, for example when calculating scores of multi-nucleotide substitutions or new InDels offline. In such cases, the sequence models can also be run on regular CPUs, however with substantial performance loss. While ESM derived scores are only calculated for coding parts of the genome, regulatory effects need to be scored for approximately 50x more genomic positions. Thus, we already implemented a simpler network for regulatory variants, so that we were able to build the new CADD model, pre-score variants and enable scoring of variants in a reasonable amount of time on CPUs. Consequently, there is a need for simpler models, that achieve good but not necessarily the best performance for genome-wide scoring purposes, and we hope to raise awareness within the scientific community. We note that this also has an environmental impact.

As an organismal and genome-wide model of variant effects, CADD cannot represent the significance of individual genes for specific diseases (78,79). Existing gene and transcript specific information may therefore aid variant prioritization, independent and in addition to the ranking of variants by CADD scores. For example, information about the specific phenotype (including pathways, gene interactions, or affected tissues) is potentially of high relevance. This may motivate the integration of gene or transcript level annotations into genome-wide models like CADD. However, if we use measures that are specific to each gene and transcript, such as essentiality, protein interactions and network centrality, or expression specificity, this could in principle also impair the discovery of less well-studied disease genes due to observation biases (80). Further, to combine annotations into genome-wide models, they should be at nucleotide resolution, available for all variants in a class, and without major biases. Thus, even though other information is useful for a final variant ranking, we are currently skeptical of integrating broad-resolution annotations that prioritize variants based on their location in specific genomic regions.

Another extension of CADD that we have been considering is the extension, or better described as a split-out, to cell-type specific models. Tens of thousands of functional genomics data sets are available, for example through the NBCI Gene Expression Omnibus (GEO) (81)—including gene expression, DNA accessibility, immunoprecipitation of DNA binding (transcription factors and histones), DNA methylation, 3D organization and interaction of DNA elements (e.g. enhancer-promoter links) for various cell-types, whole tissues or single cell experiments. The ENCODE project and others initiated data portals with versioned and uniform data processing pipelines as well as explored imputation of molecular assays resulting from such data (3,82). Further, the availability of single cell atlas data has substantially increased over the last years. While an individual integration is clearly not feasible, it can be speculated that cell-type specific sequence model representation of these data sets could reduce the sheer number of data sets and has the potential to remove individual biases by combining many experiments. However, cell-types are not equidistant in expression and molecular profiles. Thus, without a trajectory or 'phylogenetic' tree of cell-types, the result would be many correlated CADD models for some selection of cell-types. It is therefore critical to first describe the relationship between cell-types and to integrate data across a continuous Waddington-like landscape of cell-types and cell states (83). Such a model would be the goal for bringing cell-type effects to the challenge of identifying disease or phenotype causal variants on the organismal level.

On an unrelated note, CADD v1.7 might be the last version to support the human reference genome sequence GRCh37, a version that was succeeded by the better and more complete GRCh38 version in December 2013. In the future, we want to focus on the support of GRCh38 as well as more complete representations of human genetic sequence and variation. GRCh38 and recent efforts by the Telomere-to-Telomere (T2T) consortium have clearly shown the limitations of a single reference genome (84). The current T2T assembly, while more comprehensive than GRCh38, also misses genomic segments of diverse human genomes. Pangenome and graph representation efforts (85–87) will provide an interesting framework, with major challenges for the mapping and representation of existing gene and transcript annotations, sequence conservation, or biochemical readouts (e.g. ENCODE ChIP and DNase data). With only initial mappings of coordinates between pangenomes and GRCh38 being developed, it will be a major effort to develop unbiased CADD scores in such settings.

## Data availability

The training and validation data underlying this article are available at https://cadd.bihealth.org/training. Pre-scored whole genome SNV files are available at https://cadd.bihealth.org/download.

## Supplementary data

Supplementary Data are available at NAR Online.

## Acknowledgements

We thank current and previous members of the Kircher lab for helpful discussions and suggestions. Specifically, we would like to acknowledge Pyaree Mohan Dash for computation of the Enformer scores. Computation has been performed on the HPC for Research cluster of the Berlin Institute of Health at Charité – Universitätsmedizin Berlin as well as on the OMICS cluster at the University of Lübeck.

## Funding

Berlin Institute of Health at Charité – Universitätsmedizin Berlin (to M.K., M.S.); Deutsche Forschungsgemeinschaft (DFG) [464313370 to M.S.]; Impact of Genomic Variation on Function/NHGRI [1UM1HG011966-01 to M.K., T.M.]; Helmholtz Einstein International Berlin Research School in Data Science (to M.K., L.N.); University of Lübeck (to M.K.). Funding for open access charge: Berlin Institute of Health at Charité – Universitätsmedizin Berlin.

## Conflict of interest statement

None declared.

## References

1. Shendure,J., Findlay,G.M. and Snyder,M.W. (2019) Genomic medicine – progress, pitfalls, and promise. *Cell*, **177**, 45–57.
2. Gibbs,R.A. (2020) The Human Genome Project changed everything. *Nat. Rev. Genet.*, **21**, 575–576.
3. Moore,J.E., Purcaro,M.J., Pratt,H.E., Epstein,C.B., Shoresh,N., Adrian,J., Kawli,T., Davis,C.A., Dobin,A., Kaul,R., *et al.* (2020) Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature*, **583**, 699–710.
4. Gurdasani,D., Barroso,I., Zeggini,E. and Sandhu,M.S. (2019) Genomics of disease risk in globally diverse populations. *Nat. Rev. Genet.*, **20**, 520–535.
5. Claussnitzer,M., Cho,J.H., Collins,R., Cox,N.J., Dermitzakis,E.T., Hurles,M.E., Kathiresan,S., Kenny,E.E., Lindgren,C.M., MacArthur,D.G., *et al.* (2020) A brief history of human disease genetics. *Nature*, **577**, 179–189.
6. Kingdom,R. and Wright,C.F. (2022) Incomplete penetrance and variable expressivity: from clinical studies to population cohorts. *Front. Genet.*, **13**, 920390.
7. Chatterjee,S. and Ahituv,N. (2017) Gene regulatory elements, major drivers of Human disease. *Annu. Rev. Genomics Hum. Genet.*, **18**, 45–63.
8. Spielmann,M., Lupiáñez,D.G. and Mundlos,S. (2018) Structural variation in the 3D genome. *Nat. Rev. Genet.*, **19**, 453–467.
9. Gasperini,M., Tome,J.M. and Shendure,J. (2020) Towards a comprehensive catalogue of validated and target-linked human enhancers. *Nat. Rev. Genet.*, **21** , 292–310.
10. Przybyla,L. and Gilbert,L.A. (2022) A new era in functional genomics screens. *Nat. Rev. Genet.*, **23**, 89–103.
11. Findlay,G.M. (2021) Linking genome variants to disease: scalable approaches to test the functional impact of human mutations. *Hum. Mol. Genet.*, **30**, R187–R197.
12. 100,000 Genomes Project Pilot Investigators, Smedley,D., Smith,K.R., Martin,A., Thomas,E.A., McDonagh,E.M., Cipriani,V., Ellingford,J.M., Arno,G., Tucci,A., *et al.* (2021) 100,000 Genomes pilot on rare-disease diagnosis in health care - preliminary report. *N. Engl. J. Med.*, **385**, 1868–1880.
13. Spielmann,M. and Kircher,M. (2022) Computational and experimental methods for classifying variants of unknown clinical significance. *Cold Spring Harb. Mol. Case Stud.*, **8**, a006196.
14. Esposito,D., Weile,J., Shendure,J., Starita,L.M., Papenfuss,A.T., Roth,F.P., Fowler,D.M. and Rubin,A.F. (2019) MaveDB: an open-source platform to distribute and interpret data from multiplexed assays of variant effect. *Genome Biol.*, **20**, 223.
15. Shefchek,K.A., Harris,N.L., Gargano,M., Matentzoglu,N., Unni,D., Brush,M., Keith,D., Conlin,T., Vasilevsky,N., Zhang,X.A., *et al.* (2020) The Monarch Initiative in 2019: an integrative data and analytic platform connecting phenotypes to genotypes across species. *Nucleic Acids Res.*, **48**, D704–D715.
16. Ellingford,J.M., Ahn,J.W., Bagnall,R.D., Baralle,D., Barton,S., Campbell,C., Downes,K., Ellard,S., Duff-Farrier,C., FitzPatrick,D.R., *et al.* (2022) Recommendations for clinical interpretation of variants found in non-coding regions of the genome. *Genome Med*, **14**, 73.
17. Kircher,M. and Ludwig,K.U. (2022) Systematic assays and resources for the functional annotation of non-coding variants. *Med. Genet.*, **34**, 275–286.
18. Erlich,Y. and Narayanan,A. (2014) Routes for breaching and protecting genetic privacy. *Nat. Rev. Genet.*, **15**, 409–421.
19. Hudson,M., Garrison,N.A., Sterling,R., Caron,N.R., Fox,K., Yracheta,J., Anderson,J., Wilcox,P., Arbour,L., Brown,A., *et al.* (2020) Rights, interests and expectations: indigenous perspectives on unrestricted access to genomic data. *Nat. Rev. Genet.*, **21**, 377–384.
20. Kircher,M., Witten,D.M., Jain,P., O'Roak,B.J., Cooper,G.M. and Shendure,J. (2014) A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.*, **46**, 310–315.
21. Niroula,A. and Vihinen,M. (2016) Variation interpretation predictors: principles, types, performance, and choice. *Hum. Mutat.*, **37**, 579–597.
22. McInnes,G., Sharo,A.G., Koleske,M.L., Brown,J.E.H., Norstad,M., Adhikari,A.N., Wang,S., Brenner,S.E., Halpern,J., Koenig,B.A., *et al.* (2021) Opportunities and challenges for the computational interpretation of rare variation in clinically important genes. *Am. J. Hum. Genet.*, **108**, 535–548.
23. Smail,C., Ferraro,N.M., Hui,Q., Durrant,M.G., Aguirre,M., Tanigawa,Y., Keever-Keigher,M.R., Rao,A.S., Justesen,J.M., Li,X., *et al.* (2022) Integration of rare expression outlier-associated variants improves polygenic risk prediction. *Am. J. Hum. Genet.*, **109**, 1055–1064.
24. Groß,C., de Ridder,D. and Reinders,M. (2018) Predicting variant deleteriousness in non-human species: applying the CADD approach in mouse. *BMC Bioinf.*, **19**, 373.
25. Groß,C., Derks,M., Megens,H.-J., Bosse,M., Groenen,M.A.M., Reinders,M. and de Ridder,D. (2020) pCADD: SNV prioritisation in Sus scrofa. *Genet. Sel. Evol.*, **52**, 4.

26. Groß,C., Bortoluzzi,C., de Ridder,D., Megens,H.-J., Groenen,M.A.M., Reinders,M. and Bosse,M. (2020) Prioritizing sequence variants in conserved non-coding elements in the chicken genome using chCADD. *PLos Genet.*, **16**, e1009027.
27. Racimo,F. and Schraiber,J.G. (2014) Approximation to the distribution of fitness effects across functional categories in Human segregating polymorphisms. *PLoS Genet.*, **10**, e1004697.
28. Murphy,D.A., Elyashiv,E., Amster,G. and Sella,G. (2023) Broad-scale variation in human genetic diversity levels is predicted by purifying selection on coding and non-coding elements. *Elife*, **12**, e76065.
29. Kleinert,P. and Kircher,M. (2022) A framework to score the effects of structural variants in health and disease. *Genome Res.*, **32**, 766–777.
30. Sundaram,L., Gao,H., Padigepati,S.R., McRae,J.F., Li,Y., Kosmicki,J.A., Fritzilas,N., Hakenberg,J., Dutta,A., Shon,J., *et al.* (2018) Predicting the clinical impact of human mutation with deep neural networks. *Nat. Genet.*, **50**, 1161–1170.
31. Gao,H., Hamp,T., Ede,J., Schraiber,J.G., McRae,J., Singer-Berk,M., Yang,Y., Dietrich,A.S.D., Fiziev,P.P., Kuderna,L.F.K., *et al.* (2023) The landscape of tolerated genetic variation in humans and primates. *Science*, **380**, eabn8153.
32. Rentzsch,P., Schubach,M., Shendure,J. and Kircher,M. (2021) CADD-splice-improving genome-wide variant effect prediction using deep learning-derived splice scores. *Genome Med*, **13**, 31.
33. Meier,J., Rao,R., Verkuil,R., Liu,J., Sercu,T. and Rives,A. (2021) Language models enable zero-shot prediction of the effects of mutations on protein function. In: *Advances in Neural Information Processing Systems*. Curran Associates, Inc., Vol. **34**, pp. 29287–29303.
34. Christmas,M.J., Kaplow,I.M., Genereux,D.P., Dong,M.X., Hughes,G.M., Li,X., Sullivan,P.F., Hindle,A.G., Andrews,G., Armstrong,J.C., *et al.* (2023) Evolutionary constraint and innovation across hundreds of placental mammals. *Science*, **380**, eabn3943.
35. Linder,J., Koplik,S.E., Kundaje,A. and Seelig,G. (2022) Deciphering the impact of genetic variation on human polyadenylation using APARENT2. *Genome Biol.*, **23**, 232.
36. Seplyarskiy,V., Lee,D.J., Koch,E.M., Lichtman,J.S., Luan,H.H. and Sunyaev,R. (2022) A mutation rate model at the basepair resolution identifies the mutagenic effect of Polymerase III transcription. bioRxiv doi: https://doi.org/10.1101/2022.08.20.504670, 21 August 2022, preprint: not peer reviewed.
37. Landrum,M.J., Lee,J.M., Riley,G.R., Jang,W., Rubinstein,W.S., Church,D.M. and Maglott,D.R. (2014) ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.*, **42**, D980–D985.
38. Landrum,M.J., Chitipiralla,S., Brown,G.R., Chen,C., Gu,B., Hart,J., Hoffman,D., Jang,W., Kaur,K., Liu,C., *et al.* (2020) ClinVar: improvements to accessing data. *Nucleic Acids Res.*, **48**, D835–D844.
39. Katsonis,P., Wilhelm,K., Williams,A. and Lichtarge,O. (2022) Genome interpretation using in silico predictors of variant impact. *Hum. Genet.*, **141**, 1549–1577.
40. Livesey,B.J. and Marsh,J.A. (2022) Interpreting protein variant effects with computational predictors and deep mutational scanning. *Dis. Model. Mech.*, **15**, dmm049510.
41. Herrero,J., Muffato,M., Beal,K., Fitzgerald,S., Gordon,L., Pignatelli,M., Vilella,A.J., Searle,S.M.J., Amode,R., Brent,S., *et al.* (2016) Ensembl comparative genomics resources. *Database (Oxford)*, **2016**, bav096.
42. Rentzsch,P., Witten,D., Cooper,G.M., Shendure,J. and Kircher,M. (2019) CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.*, **47**, D886–D894.
43. Vitsios,D., Dhindsa,R.S., Middleton,L., Gussow,A.B. and Petrovski,S. (2021) Prioritizing non-coding regions based on human genomic constraint and sequence context with deep learning. *Nat. Commun.*, **12**, 1504.
44. Bogard,N., Linder,J., Rosenberg,A.B. and Seelig,G. (2019) A deep neural network for predicting and engineering alternative polyadenylation. *Cell*, **178**, 91–106.e23.
45. Andrews,G., Fan,K., Pratt,H.E., Phalke,N., Zoonomia Consortium, Karlsson,E.K., Lindblad-Toh,K., Gazal,S., Moore,J.E. and Weng,Z. (2023) Mammalian evolution of human cis-regulatory elements and transcription factor binding sites. *Science*, **380**, eabn7930.
46. Livesey,B.J. and Marsh,J.A. (2023) Updated benchmarking of variant effect predictors using deep mutational scanning. *Mol. Syst. Biol.*, **19**, e11474.
47. Brandes,N., Goldman,G., Wang,C.H., Ye,C.J. and Ntranos,V. (2023) Genome-wide prediction of disease variant effects with a deep protein language model. *Nat. Genet.*, **55**, 1512–1522.
48. Avsec,Ž., Agarwal,V., Visentin,D., Ledsam,J.R., Grabska-Barwinska,A., Taylor,K.R., Assael,Y., Jumper,J., Kohli,P. and Kelley,D.R. (2021) Effective gene expression prediction from sequence by integrating long-range interactions. *Nat. Methods*, **18**, 1196–1203.
49. Inoue,F., Kircher,M., Martin,B., Cooper,G.M., Witten,D.M., McManus,M.T., Ahituv,N. and Shendure,J. (2017) A systematic comparison reveals substantial differences in chromosomal versus episomal encoding of enhancer activity. *Genome Res.*, **27**, 38–52.
50. Shigaki,D., Adato,O., Adhikari,A.N., Dong,S., Hawkins-Hooker,A., Inoue,F., Juven-Gershon,T., Kenlay,H., Martin,B., Patra,A., *et al.* (2019) Integration of multiple epigenomic marks improves prediction of variant impact in saturation mutagenesis reporter assay. *Hum. Mutat.*, **40**, 1280–1291.
51. Andersson,R. and Sandelin,A. (2020) Determinants of enhancer and promoter activities of regulatory elements. *Nat. Rev. Genet.*, **21**, 71–87.
52. Notin,P., Dias,M., Frazer,J., Hurtado,J.M., Gomez,A.N., Marks,D. and Gal,Y. (2022) Tranception: protein fitness prediction with autoregressive transformers and inference-time retrieval. In: *Proceedings of the 39th International Conference on Machine Learning*. PMLR. Vol. **162**, pp. 16990–17017.
53. Cannon,S., Williams,M., Gunning,A.C. and Wright,C.F. (2023) Evaluation of in silico pathogenicity prediction tools for the classification of small in-frame indels. *BMC Med. Genet.*, **16**, 36.
54. Karczewski,K.J., Francioli,L.C., Tiao,G., Cummings,B.B., Alföldi,J., Wang,Q., Collins,R.L., Laricchia,K.M., Ganna,A., Birnbaum,D.P., *et al.* (2020) The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, **581**, 434–443.
55. Wright,C.F., Fitzgerald,T.W., Jones,W.D., Clayton,S., McRae,J.F., van Kogelenberg,M., King,D.A., Ambridge,K., Barrett,D.M., Bayzetinova,T., *et al.* (2015) Genetic diagnosis of developmental disorders in the DDD study: a scalable analysis of genome-wide research data. *Lancet*, **385**, 1305–1314.
56. Krützfeldt,L.-M., Schubach,M. and Kircher,M. (2020) The impact of different negative training data on regulatory sequence predictions. *PLoS One*, **15**, e0237412.
57. Kircher,M., Xiong,C., Martin,B., Schubach,M., Inoue,F., Bell,R.J.A., Costello,J.F., Shendure,J. and Ahituv,N. (2019) Saturation mutagenesis of twenty disease-associated regulatory elements at single base-pair resolution. *Nat. Commun.*, **10**, 3583.
58. Karollus,A., Mauermeier,T. and Gagneur,J. (2023) Current sequence-based models capture gene expression determinants in promoters but mostly ignore distal enhancers. *Genome Biol.*, **24**, 56.
59. Agarwal,V., Inoue,F., Schubach,M., Martin,B.K., Dash,P.M., Zhang,Z., Sohota,A., Noble,W.S., Yardimci,G.G., Kircher,M., *et al.* (2023) Massively parallel characterization of transcriptional regulatory elements in three diverse human cell types. bioRxiv doi: https://doi.org/10.1101/2023.03.05.531189, 06 March 2023, preprint: not peer reviewed.
60. Mather,C.A., Mooney,S.D., Salipante,S.J., Scroggins,S., Wu,D., Pritchard,C.C. and Shirts,B.H. (2016) CADD score has limited

clinical validity for the identification of pathogenic variants in noncoding regions in a hereditary cancer panel. *Genet. Med.*, **18**, 1269–1275.

61. Wang,D., Li,J., Wang,Y. and Wang,E. (2022) A comparison on predicting functional impact of genomic variants. *NAR Genom. Bioinform.*, **4**, lqab122.

62. Schubach,M., Nazaretyan,L. and Kircher,M. (2022) The regulatory mendelian mutation score for GRCh38. *Gigascience*, **12**, giad024.

63. Schmidt,A., Röner,S., Mai,K., Klinkhammer,H., Kircher,M. and Ludwig,K.U. (2023) Predicting the pathogenicity of missense variants using features derived from AlphaFold2. *Bioinformatics*, **39**, btad280.

64. The 1000 Genomes Project Consortium, Abecasis,G.R., Auton,A., Brooks,L.D., DePristo,M.A., Durbin,R.M., Handsaker,R.E., Kang,H.M., Marth,G.T. and McVean,G.A. (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature*, **491**, 56–65.

65. Martin,F.J., Amode,M.R., Aneja,A., Austine-Orimoloye,O., Azov,A.G., Barnes,I., Becker,A., Bennett,R., Berry,A., Bhai,J., *et al.* (2023) Ensembl 2023. *Nucleic Acids Res.*, **51**, D933–D941.

66. McLaren,W., Gil,L., Hunt,S.E., Riat,H.S., Ritchie,G.R.S., Thormann,A., Flicek,P. and Cunningham,F. (2016) The Ensembl variant effect predictor. *Genome Biol.*, **17**, 122.

67. Sharo,A.G., Zou,Y., Adhikari,A.N. and Brenner,S.E. (2023) ClinVar and HGMD genomic variant classification accuracy has improved over time, as measured by implied disease burden. *Genome Med*, **15**, 51.

68. van der Sanden,B.P.G.H., Schobers,G., Corominas Galbany,J., Koolen,D.A., Sinnema,M., van Reeuwijk,J., Stumpel,C.T.R.M., Kleefstra,T., de Vries,B.B.A., Ruiterkamp-Versteeg,M., *et al.* (2023) The performance of genome sequencing as a first-tier test for neurodevelopmental disorders. *Eur. J. Hum. Genet.*, **31**, 81–88.

69. Liu,X., Li,C., Mou,C., Dong,Y. and Tu,Y. (2020) dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs. *Genome Medicine*, **12**, 103.

70. Wang,K., Li,M. and Hakonarson,H. (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.*, **38**, e164–e164.

71. Ng,S.B., Turner,E.H., Robertson,P.D., Flygare,S.D., Bigham,A.W., Lee,C., Shaffer,T., Wong,M., Bhattacharjee,A., Eichler,E.E., *et al.* (2009) Targeted capture and massively parallel sequencing of 12 human exomes. *Nature*, **461**, 272–276.

72. Chen,S., Francioli,L.C., Goodrich,J.K., Collins,R.L., Kanai,M., Wang,Q., Alföldi,J., Watts,N.A., Vittal,C., Gauthier,L.D., *et al.* (2022) A genome-wide mutational constraint map quantified from variation in 76,156 human genomes. bioRxiv doi: https://doi.org/10.1101/2022.03.20.485034, 21 March 2022, preprint: not peer reviewed.

73. Zhang,P., Bigio,B., Rapaport,F., Zhang,S.-Y., Casanova,J.-L., Abel,L., Boisson,B. and Itan,Y. (2018) PopViz: a webserver for visualizing minor allele frequencies and damage prediction scores of human genetic variations. *Bioinformatics*, **34**, 4307–4309.

74. Bonomi,L., Huang,Y. and Ohno-Machado,L. (2020) Privacy challenges and research opportunities for genomic data sharing. *Nat. Genet.*, **52**, 646–654.

75. Mölder,F., Jablonski,K.P., Letcher,B., Hall,M.B., Tomkins-Tinch,C.H., Sochat,V., Forster,J., Lee,S., Twardziok,S.O., Kanitz,A., *et al.* (2021) Sustainable data analysis with Snakemake. *F1000Res*, **10**, 33.

76. King,M.C. and Wilson,A.C. (1975) Evolution at two levels in humans and chimpanzees. *Science*, **188**, 107–116.

77. Kuderna,L.F.K., Gao,H., Janiak,M.C., Kuhlwilm,M., Orkin,J.D., Bataillon,T., Manu,S., Valenzuela,A., Bergman,J., Rousselle,M., *et al.* (2023) A global catalog of whole-genome diversity from 233 primate species. *Science*, **380**, 906–913.

78. Havrilla,J.M., Pedersen,B.S., Layer,R.M. and Quinlan,A.R. (2019) A map of constrained coding regions in the human genome. *Nat. Genet.*, **51**, 88.

79. Abramovs,N., Brass,A. and Tassabehji,M. (2020) GeVIR is a continuous gene-level metric that uses variant distribution patterns to prioritize disease candidate genes. *Nat. Genet.*, **52**, 35–39.

80. Stoeger,T., Gerlach,M., Morimoto,R.I. and Amaral,L.A.N. (2018) Large-scale investigation of the reasons why potentially important genes are ignored. *PLoS Biol.*, **16**, e2006643.

81. Barrett,T., Wilhite,S.E., Ledoux,P., Evangelista,C., Kim,I.F., Tomashevsky,M., Marshall,K.A., Phillippy,K.H., Sherman,P.M., Holko,M., *et al.* (2013) NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res.*, **41**, D991–D995.

82. Ernst,J. and Kellis,M. (2017) Chromatin-state discovery and genome annotation with ChromHMM. *Nat. Protoc.*, **12**, 2478–2492.

83. Domcke,S. and Shendure,J. (2023) A reference cell tree will serve science better than a reference cell atlas. *Cell*, **186**, 1103–1114.

84. Nurk,S., Koren,S., Rhie,A., Rautiainen,M., Bzikadze,A.V., Mikheenko,A., Vollger,M.R., Altemose,N., Uralsky,L., Gershman,A., *et al.* (2022) The complete sequence of a human genome. *Science*, **376**, 44–53.

85. Computational Pan-Genomics Consortium (2018) Computational pan-genomics: status, promises and challenges. *Brief Bioinform*, **19**, 118–135.

86. Eizenga,J.M., Novak,A.M., Sibbesen,J.A., Heumos,S., Ghaffaari,A., Hickey,G., Chang,X., Seaman,J.D., Rounthwaite,R., Ebler,J., *et al.* (2020) Pangenome graphs. *Annu. Rev. Genomics Hum. Genet.*, **21**, 139–162.

87. Liao,W.-W., Asri,M., Ebler,J., Doerr,D., Haukness,M., Hickey,G., Lu,S., Lucas,J.K., Monlong,J., Abel,H.J., *et al.* (2023) A draft human pangenome reference. *Nature*, **617**, 312–324.

**Received:** September 14, 2023. **Revised:** October 14, 2023. **Editorial Decision:** October 16, 2023. **Accepted:** October 17, 2023