CAGI7 Conference

Accepted Abstracts

Northeastern University

December 6-8, 2025

Evaluation of In-Frame Indel Predictors using Saturation Genome Editing (SGE) Data

Haneen Abderrazzaq[1], Mugdha Singh[2,3], Silvia Casadei[4,5], Matthew Snyder[4,5], Nahum Smith[4,5], Lea Starita[4,5], Anne O'Donnell-Luria[2,3], Predrag Radivojac[1,*]

[1]Khoury College of Computer Sciences, Northeastern University, Boston, Massachusetts, USA
[2]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard; Cambridge, Massachusetts, USA
[3]Division of Genetics and Genomics, Boston Children's Hospital, Harvard Medical School, Boston, Massachusetts, USA
[4]Department of Genome Sciences, University of Washington, Seattle, Washington, USA
[5]Brotman Baty Institute for Precision Medicine, Seattle, Washington, USA

*Corresponding author: Predrag Radivojac (predrag@northeastern.edu)

Insertions and deletions (indels) are associated with diverse functional consequences. However, non-frameshifting indels remain understudied, largely due to limited data availability in clinical databases. This data scarcity has hindered the development and assessment of indel-specific variant effect predictors. Saturation genome editing (SGE) addresses this challenge by measuring the functional impact of all possible variants within target genes. In this work, we leverage SGE data from six clinically relevant genes (BARD1, CTCF, PALB2, RAD51D, SFPQ, and XRCC2), to evaluate the performance of in-frame indel variant effect predictors. We assess a diverse set of computational tools, including traditional machine learning models and newer protein language models, on every possible three base pair (single amino acid) deletion for each gene. Our findings demonstrate how integrating computational models with large-scale functional data can improve our understanding of indel variant effects.

Annotation database updates and their impact on missense variant effect predictions

Timothy Bergquist[1], Vikas Pejaver*,[1,2]
[1]Institute for Genomic Health, Icahn School of Medicine at Mount Sinai
[2]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai
*vikas.pejaver@mssm.edu

Recent work has suggested that computational tools that infer the functional impact of genetic variants (e.g., AlphaMissense, MutPred2, REVEL) can provide stronger evidence towards the assertion of variant pathogenicity/benignity than previously stipulated by the American College of Genetics and Genomics (ACMG) and the Association for Molecular Pathology (AMP) guidelines for clinical variant classification. For scalability and convenience, it is common practice for missense variant impact predictor developers to precompute scores for all theoretically possible variants in the human exome and make them available to end-users, often through annotation resources (e.g., VEP). However, these resources are continually updated with the most current genetic and molecular information, involving changes to the mappings between nucleotide coordinates and amino acid positions, the protein sequence, the designation of “canonical” status to an isoform, among others. Thus, precomputed scores may not necessarily be in synchrony with updates to bioinformatics databases and may be discordant if not generated using the latest data. Here, we quantify the extent to which this influences clinical variant classification by systematically predicting impact on the same variants across multiple releases of the Ensembl database.

We found that of 269,143 ClinVar missense variants from clinically actionable genes in ClinGen, 36,378 (~13%) were affected by changes in Ensembl between releases 90 and 115. Changes included a shift in canonical designation, protein sequence changes, and changes to coding/non-coding designation. Furthermore, when MutPred2 predictions on those variants were grouped into evidential strength categories as per the ACMG/AMP guidelines, database changes often shifted the strength of evidence of the affected variants from the evidence strength *Indeterminate* to *Supporting* pathogenicity/benignity and vice versa with 16,734 (~6%) variants of the 269,143 missense variants changing in evidential strength. As future ACMG/AMP guidelines envision a more prominent role for variant impact predictors in clinical variant classification, our results have direct implications for their proper incorporation into clinical workflows.

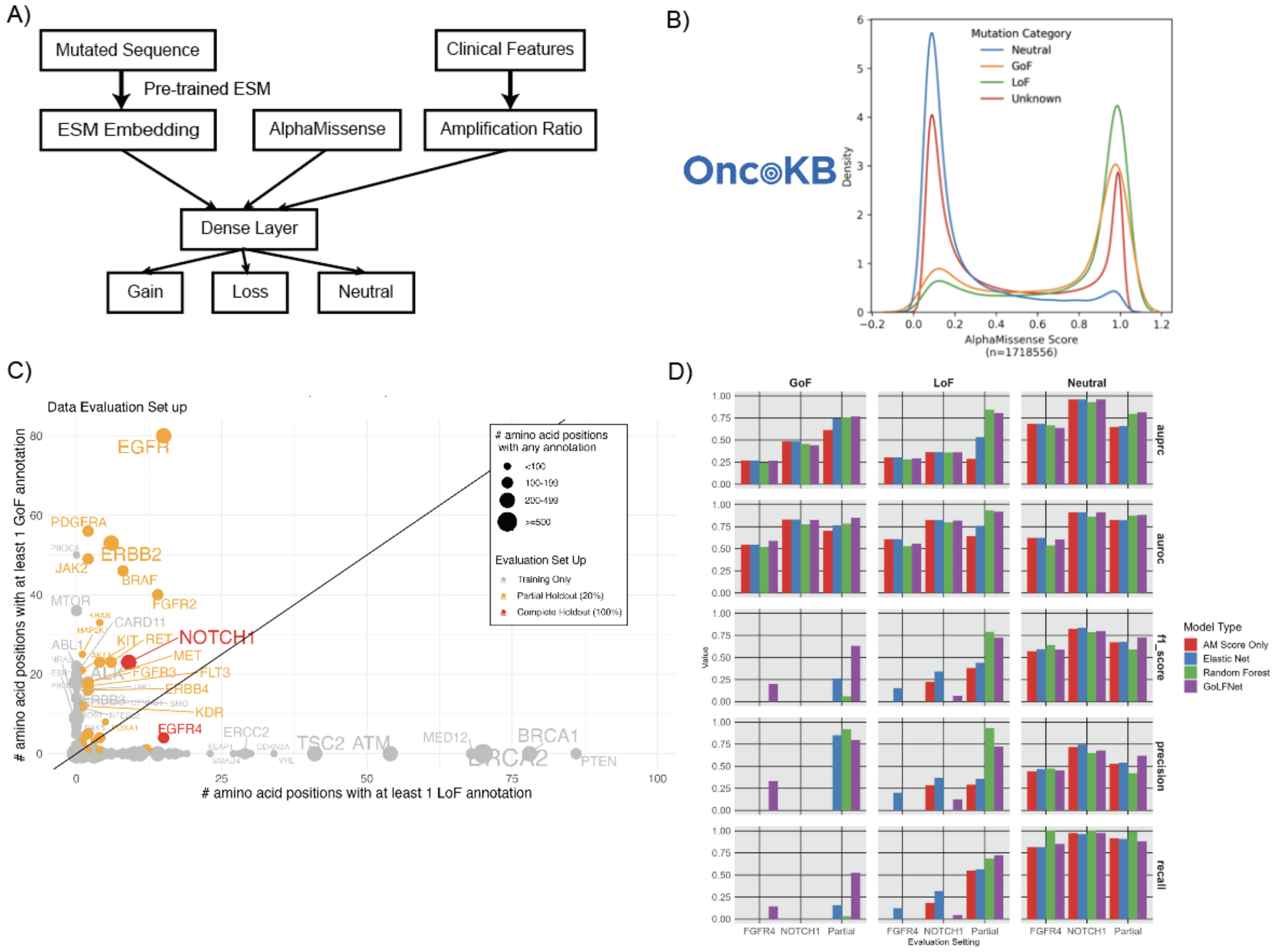
Functional Prediction of Somatic Missense Variants using Large Protein Language Model

Hoyin Chu[1,*], Caleb Lareau[1,*]

1. Computational & Systems Biology Program, Memorial Sloan Kettering Cancer Center

Email: chuh@mskcc.org, lareauc@mskcc.org

The functional consequences of most somatic missense mutations in cancer remain unknown. To address this, we developed GoLFNet, a neural network trained on protein representations from large protein language model and cancer-specific features (panel A, B), to predict the functional impact of missense mutations in a large cohort of cancer patients using OncoKB labels. Compared to baseline methods, GoLFNet had superior performance in predicting the functional consequence of known missense variants both in proteins with somatic mutations included in the training set as well as in proteins out of the training set (panel C, D). In addition, we applied GoLFNet to 9460 recurrent somatic mutations with no existing annotations and identified several candidate gain-of-function mutations in BTK for further functional characterization. The model and the predicted scores for the missense mutations are available in the GitHub link: https://github.com/hoyinchu/GoLFNet



EvoStructCLIP: A Generalizable Multimodal Framework Applied Across CAGI7 Challenges

Kyungkeon Chung, Jaekyung Lee, ChungHeorn Lee, Junyoung Park, Hyejin Lee*

UXFactory Inc., Republic of Korea

*Corresponding author: hyejin@uxf.ai, Presenting author: kunnyjung@uxf.ai

Accurate interpretation of missense variants requires models capable of integrating residue-level structural constraints with long-range evolutionary dependencies. We present EvoStructCLIP, a multimodal deep learning framework that jointly encodes structural and multiple sequence alignment (MSA) features through CLIP-style contrastive pretraining. The framework aligns a voxel-based 3D encoder built on MBConv3D blocks with 3D SE attention and an MSA encoder employing cross-axial Mamba blocks for long-range evolutionary modeling, generating transferable embeddings for variant effect prediction.

EvoStructCLIP was systematically applied to seven CAGI7 challenges—Annotate All Missense, ARSA, ATP7B, BARD1, FGFR, LPL, and TSC2—using a shared pretrained backbone and task-specific output heads. For the large-scale Annotate All Missense task, the MSA branch alone achieved strong generalization (PR-AUC 0.935, ROC-AUC 0.958 on ClinVar) while predicting over 70 million missense variants from dbNSFP. In quantitative assays assessing protein stability, RNA abundance, or cellular function, pretrained embeddings were concatenated with 19 stability-oriented structural and evolutionary features to form 275-dimensional representations, modeled through cross-validated ensembles of XGBoost, Random Forest, Gradient Boosting, and MLP regressors. Scaling strategies—including temperature, logistic, Min–Max, and quantile normalization—were adapted to each assay to align predicted outputs with experimental distributions.

Across all tasks, EvoStructCLIP demonstrated consistent performance without task-specific retraining, underscoring the transferability of contrastively learned embeddings across pathogenicity, stability, and activity predictions. These results highlight the potential of multimodal contrastive pretraining as a scalable foundation for the unified interpretation of variant effects across structural and evolutionary dimensions.

The CAGI7 Clinical Genomes Challenge: pathogenic variant identification in rare disease patients from the Rare Genomes Project

Laura E Covill[1,2,*], Stephanie DiTroia[1], Melanie O'Leary[1], Predrag Radivojac[3], Heidi Rehm[1,4], Anne O'Donnell-Luria[1,2]

[1]Broad Institute of MIT and Harvard, Cambridge, MA, USA
[2]Division of Genetics and Genomics, Boston Children's Hospital, Harvard Medical School, Boston, MA, USA
[3]Northeastern University, Boston, MA, USA
[4]Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA
*Corresponding author (email address: lcovill@broadinstitute.org)

Molecular diagnosis remains an obstacle to optimal care in rare disease, with a confirmed diagnosis only reached for 50% of patients. Genome sequencing has the potential to provide answers for an increasing number of patients, but cutting-edge annotation and prioritization tools are required to supply useful and expedient information to clinical labs and care providers. To test the efficacy of such tools in a real-world setting, we have launched the Clinical Genomes challenge as part of the 7th iteration of the Critical Assessment of Genome Interpretation (CAGI7).

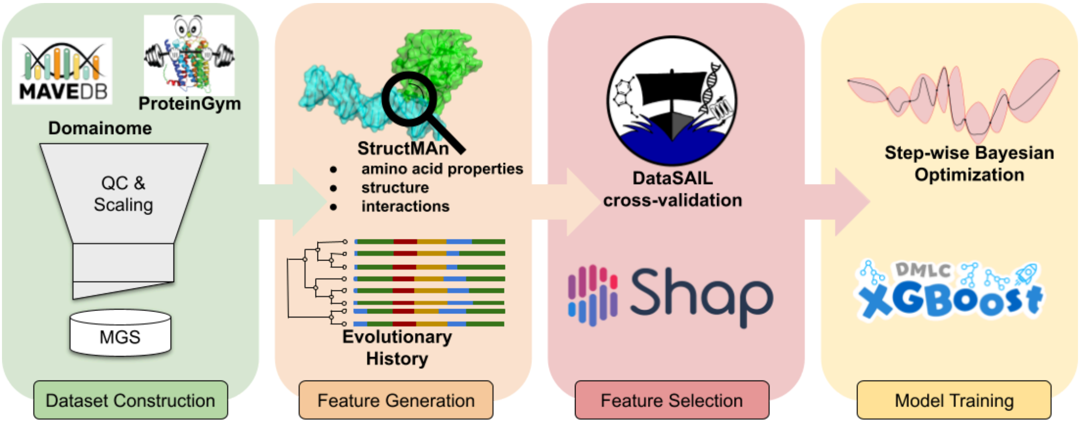
Here, we present the structure and goals of the CAGI7 Clinical Genomes challenge. Teams are challenged to use their model to rank and score variants from rare disease patient genomes and phenotypes collected as part of the Rare Genomes Project (RGP). The Clinical Genomes challenge involves 50 RGP families split into a Test Set (30 solved and unsolved families), and a Discovery Set (20 unsolved families). Up to 100 top-ranked variants per case will be submitted by challenge participants. Results for solved cases will be used to evaluate the efficacy and reliability of each model's predictions. For unsolved cases in the Test and Discovery sets, an analyst will evaluate the results of the top-performing models to identify any promising candidates for downstream analyses. Furthermore, we aim to collect data on general platform usability and services such as filtering, reporting, and reanalysis. These results have the potential to inform future model development, and platform selection by diagnostic clinical labs.

StructGuy: Data leakage free prediction of functional effects of genetic variants.

Alexander Gress*[1], Carene Benasolo[1], Johanna Becher[1], Dominique Mias-Lucquin[1], Roman Joeres[1], Sebastian Keller[1], Olga V. Kalinina[1,2,3]

[1]Helmholtz Institute for Pharmaceutical Research Saarland (HIPS), Helmholtz Centre for Infection Research (HZI), Saarbruecken, Germany
[2]Center for Bioinformatics, Saarland University, Saarbruecken, Germany
[3]Faculty of Medicine, Saarland University, Homburg, Germany

(*) alexander.gress@helmholtz-hips.de



Variant effect prediction for protein-coding genes – predicting how genetic variants affect protein function – has drawn recently much attention of the biological machine learning community. Multiplexed assays of variant effects (MAVE) experiments serve as a rich data source, but cannot deliver enough data for training truly large neural-net models. Hence, zero-shot methods, for example protein language models, have increasingly gained popularity. For these methods, MAVE results serve primarily for evaluation purposes, as exemplified by the ProteinGym benchmark. In this study, we argue that the rapidly increasing amounts of MAVE data can be used to train efficient supervised methods, presenting our new tool StructGuy, based on gradient boosting trees methodology. In contrast to other supervised methods in the field, StructGuy, thanks to its dedicated training dataset and data leakage-free training process, can predict variant effects for proteins not seen during training. To evaluate this generalization ability, we constructed a dedicated benchmark and compared StructGuy with zero-shot methods from the ProteinGym leaderboard achieving a competitive performance. Further, we demonstrate that thanks to its architecture and careful feature engineering, we are able to provide fully interpretable predictions and direct explanations of the influence of mutations on protein three-dimensional structure, which favourably differs StructGuy from zero-shot tools.

ClearVariantPro: Target-Specific Embedding Space Optimization with Multi-Task Learning

Yoojin Kim, Sungnam Kim, Wonseok Chung, Kyoungyeul Lee, Doyeon Ha*

[1] Department of Artificial Intelligence, Research and Development Center, 3billion, Inc., Seoul, Republic of Korea

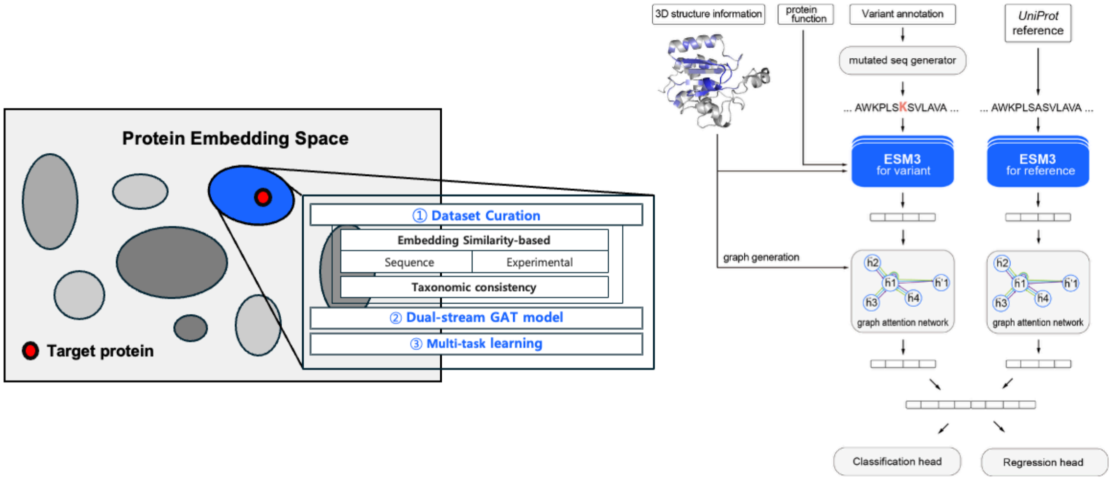
Email: dy.ha@3billion.io, yj.kim@3billion.io

Abstract

Accurate prediction of protein fitness is essential for understanding disease mechanisms and guiding protein engineering. However, this remains challenging as variant effects reside in distinct embedding spaces shaped by gene-specific and experimental contexts, hindering model generalization to unseen genes. To address this limitation, we present ClearVariantPro (CVP), a deep learning framework that optimizes target-specific embedding spaces.

CVP introduces three complementary advances spanning data curation, model architecture, and training strategy. First, a context-aware data curation selects training samples based on sequence and experimental similarity for biologically grounded representation learning. Second, a dual-stream Graph Attention Network (GAT) jointly encodes 3D structures of wild-type and mutated proteins, enabling effective modeling of functional consequences through comparative representation. Third, a Multi-Task Learning (MTL) strategy employs dual prediction heads on unified ESM3–GAT features to simultaneously address two tasks: the Activation Screen Enrichment regression, which estimates the magnitude of functional impact, and the Activation Probability Classification, which predicts 3-class categorical activation states.

CVP achieves a Spearman's ρ of 0.4187 on the FGFR1 test set from MaveDB, representing a 42% improvement over the zero-shot performance of the *ESM3-open* model (ρ = 0.2950). Overall, CVP demonstrates that integrating biologically informed data curation, structure-aware modeling, and multi-task training enables robust variant effect prediction in zero-shot settings.



ClearVariantPro: Target-Specific Embedding Space Optimization for Zero-Shot Protein Fitness Prediction

Yoojin Kim, Sungnam Kim, Wonseok Chung, Kyoungyeul Lee, Doyeon Ha*

[1] Department of Artificial Intelligence, Research and Development Center, 3billion, Inc., Seoul, Republic of Korea

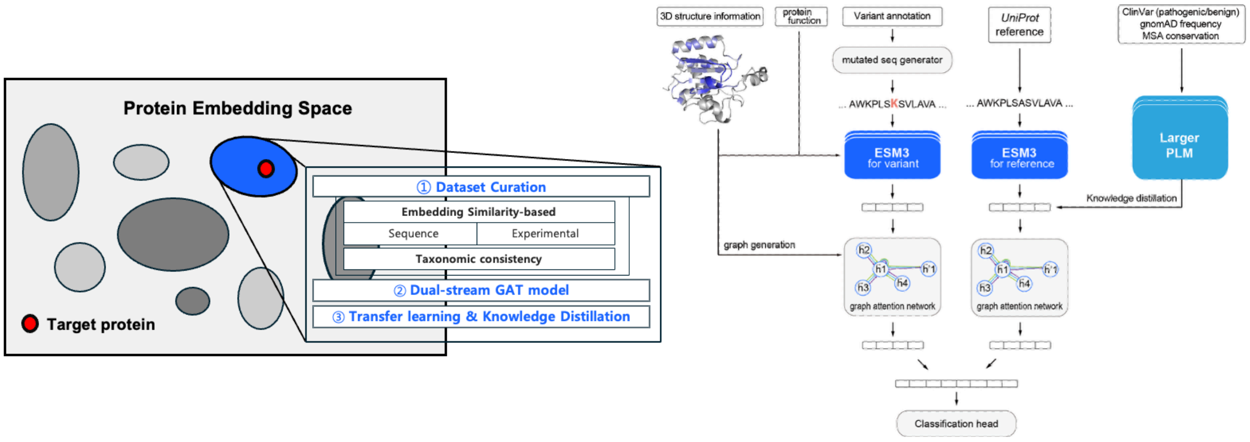
Email: dy.ha@3billion.io, yj.kim@3billion.io

Abstract

Accurate prediction of protein fitness is essential for understanding disease mechanisms and guiding protein engineering. However, this remains challenging as variant effects reside in distinct embedding spaces shaped by gene-specific and experimental contexts, hindering model generalization to unseen genes. To address this limitation, we present ClearVariantPro (CVP), a deep learning framework that optimizes target-specific embedding spaces.

CVP introduces three complementary advances spanning data curation, model architecture, and training strategy. First, a context-aware data curation selects training samples based on sequence and experimental similarity for biologically grounded representation learning. Second, a dual-stream Graph Attention Network (GAT) jointly encodes 3D structures of wild-type and mutated proteins, enabling effective modeling of functional consequences through comparative representation. Third, transfer learning from our previous pathogenicity classifier, *ClearVariantNet*, is combined with knowledge distillation using pseudo-labels from our prior variant predictor, *3Cnet*, facilitating efficient adaptation to infer variant effects on unseen genes.

CVP achieves a mean Spearman's ρ of 0.5495 across four test genes, a 23% improvement over *3Cnet* from CAGI6 (ρ = 0.4463). Specifically, CVP was evaluated on three ProteinGym proteins similar to CAGI7 challenges (TSC2, ATP7B, LPL) and directly tested on the CAGI7 ARSA dataset, showing consistent performance across diverse proteins. Overall, CVP demonstrates that integrating biologically informed data curation, structure-aware modeling, and knowledge-guided training enables robust variant effect prediction in zero-shot settings.



Variant Impact Predictor database (VIPdb) version 3 will be enabled by automated curation assistance

Yu-Jen Lin[1], Anjali Sujithan[1], Steven E. Brenner[1, *]

[1]University of California, Berkeley

jenniferyjlin@berkeley.edu

brenner@compbio.berkeley.edu

Variant interpretation is essential for identifying patients' disease-causing genetic variants amongst the millions detected in their genomes. Hundreds of Variant Impact Predictors (VIPs), also known as Variant Effect Predictors (VEPs), have been developed, spanning diverse variant types, modeling strategies, and use cases. To facilitate the exploration of VIP options, we previously released VIPdb version 2, in which we manually curated 407 VIPs with standardized metadata on variant classes, methodological features, availability, and CAGI assessments. In VIPdb version 3, we expand the metadata to include further features such as training data, capacity to predict gain of function, ability to evaluate nonsense variants, and prediction objective (e.g., clinical pathogenicity, stability, enzyme activity, splicing regulation).

VIPdb version 3 is being constructed with an automated curation-assistance framework that enables systematic updates. First, a paper selection module proposes articles to be included in VIPdb. The module retrieves PubMed abstracts of candidate papers, converts each abstract into a feature vector, and uses a linear SVM classifier to identify candidate VIP publications. Second, a curation module proposes values for each VIPdb categorical field (e.g., training data, gain-of-function, nonsense, authors, licenses). The curation module retrieves full text articles, segments and embeds documents in a vector store, and performs question-conditioned semantic search to identify passages relevant to each categorical field. These passages are then passed to a Llama-based large language model in a retrieval-augmented generation setting, which returns proposed field assignments, confidence scores, and concise rationales, for review by curators.

We will use this framework to curate newly published VIPs and new fields into VIPdb version 3, which will be made available via the VIPdb website at https://genomeinterpretation.org/vipdb

Minigene predictions with customized SpliceAI

Wanru Lin[1], Brynja Matthiasardotti[1,2], Steve Mount[1]*

1. Dept. of Cell Biology and Molecular Genetics, University of Maryland, College Park 20742
2. National Human Genome Research Institute, National Institutes of Health, Bethesda, MD 20894
WL: wlin2345@umd.edu   SM: smount@umd.edu

Evaluating the impact of sequence variants on pre-mRNA splicing is an important problem in human molecular genetics. For many years, various mini-gene constructs have been used for assessment of the impact of sequence variants on splicing. However, different mini-gene vectors give different results (Rhine et al. 2019), presumably due to the effect of flanking exons. Furthermore, recent computational predictors may have fewer errors than minigene assays.

The splicing mini-gene challenge in CAGI7 seeks predictions of the impact of splicing variants on splicing in the context of a minigene. Because SpliceAI (Jaganathan 2019) achieves greater performance when the sequence of flanking exons are included, we approach this problem by providing SpliceAI with information about the sequence context in the mini-gene vector rather than using the predicted impact of variants in their native genomic context.

Because the influence of flanking exon and intron sequences on the impact of variants on splicing is likely to depend on characteristics of the specific exon, additional features (branch site, core splice site features, exonic splicing regulatory elements, exon length, GC content, etc.) were combined with raw SpliceAI scores in a random forest model using the training data.

Our results in the challenge will be presented, as will an analysis of the context-dependence of SpliceAI predictions.

References:

Rhine et al. 2019. "Future directions for high-throughput splicing assays in precision medicine" Hum. Mutat. 40:1225-1234. doi: 10.1002/humu.23866.

Jaganathan et al. 2019 "Predicting splicing from primary sequence with deep learning." Cell 176: 535–48.e24. doi: 10.1016/j.cell.2018.12.015

Beyond Sequence: How Structure and Multimodal Models Improve Mutation-Effect Prediction

Yuxuan Liu, Yihong Yang, Rujie Yin, Hao Zheng, Orlando Haye, Joshua Park, Kevin Nguyen, and Yang Shen*

Departments of Electrical & Computer Engineering and Computer Science & Engineering, Texas A&M University

yshen@tamu.edu

Introduction

Proteins are the fundamental machinery of life, and even a single amino-acid mutation can profoundly alter their structure and function. Accurate prediction of mutation effects is critical for understanding disease mechanisms and guiding therapeutic design. Here, we address this challenge by integrating protein sequence and structure information. Using the Deep Mutational Scanning (DMS) data from the ProteinGym dataset, we benchmark state-of-the-art protein encoders—the sequence-based language model ESM-2, the structure-based graph neural networks ESM-IF, and multimodal models such as SaProt and GearNet— to systematically compare sequence-only, structure-only, and integrated representations for supervised mutation-effect prediction. We also evaluate our structure-informed protein language model (SI-PLM) and, for the ARSA Challenge, develop ensemble machine-learning models incorporating dbNSFP-derived annotations.

Approach

We trained four multi-layer perceptron (MLP) models (3 layers each) on ProteinGym using frozen encoders as feature extractors: (1) ESM-2, (2) SaProt, (3) ESM-2 + ESM-IF + GearNet, and (4) SaProt + ESM-IF + GearNet. A 5-fold cross-validation scheme with random splits was applied, leading to five checkpoints for each model. For CAGI7 DMS tasks, we used an ensemble of all four models as the default predictor and sometimes selected checkpoints based on their pathogenicity-classification performances over target-specific mutations from ClinVar. We further assessed SI-PLM, which introduces a novel cross-modality denoising pretraining task—corrupting both sequence and structure and reconstructing them—to learn sequence–structure relationships for zero-shot mutation effect prediction. For the arylsulfatase A DMS Challenge, we further leveraged the provided labeled data and trained ensemble models using dbNSFP-derived annotations under 5-fold cross-validation, handling multiple nucleotide variants (MNVs) by aggregating the maximum scores of their component SNVs.

Results

On the ProteinGym test split, incorporating structural information consistently improved performance. While sequence-only models like ESM-2 achieved strong results (average Spearman 0.712), multimodal approaches combining sequence and structure (e.g., SaProt + ESM-IF + GearNet) delivered the highest performance (0.737). The improvement margin varied across mutation-effect types, underscoring the value of integrating structural and other biological contexts into protein language models for context-specific mutation-effect prediction.

Combining evolutionary, structural and deep learning models for superior zero-shot predictions of missense and nonsense mutations

CAGI7 Challenges: LPL, ARSA, TSC2
Team: PRESCOTT (LPL, TSC2), PRESCOTT_Sorbonne (ARSA)

Authors: Gianluca Lombardi[1], Mustafa Tekpinar[2], Alessandra Carbone[1,3,*]
[1]Sorbonne Université, CNRS, IBPS, Department of Computational, Quantitative and Synthetic Biology (CQSB), UMR7238, 75005, Paris, France
[2]Department of Physics, Faculty of Science, Van Yüzüncü Yıl University, 65080, Van, Türkiye
[3]Institut Universitaire de France, Paris, France
*alessandra.carbone@lip6.fr

Predicting the effects of missense variants on protein function and stability is essential for understanding the molecular mechanisms underlying complex diseases and improving diagnostic accuracy. In the context of the Critical Assessment of Genome Interpretation (CAGI), we addressed the challenges using different variants of ESCOTT, a computational model that combines evolutionary conservation and structural features, derived here from ColabFold alignments and AlphaFold2 predicted structures, to compute mutational scores for single-point mutations. Building on the framework of its parent model GEMME, ESCOTT scores result from a linear combination of two terms: an independent term, that captures residue conservation, and an epistatic term that quantifies the importance of mutated positions relative to the query sequence in homologous sequences carrying the same single-point mutation. In ESCOTT, the epistatic term is influenced by evolutionary conservation and structural features, specifically the physico-chemical propensities of interface residues and the circular variance that measures a residue's position within the structure. This leads to a substantially improved characterisation of mutational impact, in particular for mutations impacting protein stability, as demonstrated by the results on the ProteinGym benchmark. PRESCOTT, an extension of ESCOTT, refines these scores by applying a rank-based transformation and shift informed by allele frequencies from gnomAD. The ESCOTT model was also adapted to score nonsense mutations by modelling them as the sum of single-point deletions normalised over the entire sequence. Single deletions are scored by including the gap character in the ESCOTT alphabet and including the residue-level pLDDT from AlphaFold2 predictions. A limitation of our framework for the challenges considered is the inability to evaluate synonymous mutations, which were all assigned a default score of 1. To further improve our predictions for single-point mutations, we considered combinations of ESCOTT scores with rescaled zero-shot predictions from state-of-the-art Protein Language Models, specifically ESM2-650M, ESM-IF1, ProSST-2048 and AIDO Protein-RAG 16B, relying on classification metrics from ClinVar variants for model selection.

Finally, depending on the data available for the specific challenge, additional post-processing steps were applied to better align the ESCOTT and combined scores with the expected experimental outcomes. These adjustments modified the score ranges and distributions without altering their rank order:

- For the LPL challenge, we used ESCOTT and PRESCOTT models, with rank sorted scores, without rescaling.
- For the ARSA challenge, we combined ESCOTT and AIDO Protein-RAG 16B scores, and applied an affine transformation, inferred from the provided experimental measurements, to translate the scores in the [0, 1] interval, reflecting the remaining functional fraction of proteins at 48 hours post-expression. The same transformation was then applied to predictions for unlabelled variants.
- For the TSC2 challenge, we again considered the combination of ESCOTT and AIDO Protein-RAG 16B scores and applied CDF normalisation to project the original distribution onto a Gaussian mixture. The mixture weights were inferred by fitting a Gaussian mixture model on the original score distributions.

Extending MPRALegNet for Predicting Regulatory Variant Effects in the CAGI7 Challenge

Arif A. Rather, Yangyang Lin, Jeerthi Kannan, Dongwon Lee*

Affiliation: Boston Childrens Hospital, Boston, MA, USA; Harvard Medical School; Boston, USA; Broad Institute of MIT and Harvard; Cambridge, USA
Corresponding author: dongwon.lee@childrens.harvard.edu

Understanding how noncoding genetic variants influence gene regulation remains a central challenge in genomics. The vast majority of genetic variation occurs in noncoding regulatory regions, where functional impact is difficult to interpret. Accurately predicting the regulatory effects of such variants is essential for linking genetic variation to molecular phenotypes and disease risk. To address the CAGI7 Non-Coding Variant Interpretation Challenge (lentiMPRA), we extended the MPRALegnet framework by implementing three distinct modeling approaches. First, we developed *dLegNet method for fine-tuning for variant effect prediction.* We performed transfer learning on 90 cross-validation models using the training dataset of variant effects. The provided training data were partitioned into four balanced subsets using a cross-validation strategy, in which two subsets were used for model training, one for hyperparameter tuning, and one for testing. This 4-fold cross-validation design produced 12 distinct model configurations for each base model and fine-tuning combination, resulting in a total of 1,080 trained models. we employed a hybrid loss function combining mean squared error (MSE) and a Pearson correlation (R), defined as MSE + 0.3 x (1 – R). Final predictions were generated through model ensembling. We achieved R=0.51 with cross validation. Second, we optimized MPRALegNet with a new genomic data augmentation technique. From the initial 90 models, we selected the top 10 based on prediction of effect size in the training dataset. In a second training phase, we augmented the data by (1) incorporating ~30,000 previously excluded MPRA constructs with low barcode counts and (2) training Support Vector Regression (SVR) models with gapped k-mer features on the original MPRA dataset to identify additional genomic regions. Genome-wide SVR scoring using a 200 bp sliding window (20 bp step) yielded 1.9 million non-overlapping regions with SVR < –1 and ~3,000 regions with SVR > 1, which were used exclusively for training. This model significantly improved the prediction of MPRA activity (R=0.81 vs. original model's R=0.86). Final predictions were generated by ensembling, averaging outputs across all trained models. Third, we finetuned the optimized MPRALegNet as described in Second, using dLegNet method described in first. We made only one finetuned model for each test/validation fold pairs, resulting in 10 fine-tuned models. Final predictions were made by averaging scores across all fine-tuned models.

Annotation of Noncoding Regulatory Elements Enables Diagnosis of Rare Disease

Lindsay Romo*[1,2], Emily O'Heir[1], Melanie O'Leary[1], Nicola Whiffin[3], Heidi Rehm[1,4], Anne O'Donnell-Luria[1,2]

[1]Broad Institute of MIT and Harvard, Cambridge, MA, USA; [2]Boston Children's Hospital, Harvard Medical School, Boston, MA, USA; [3]University of Oxford, Oxford, UK; [4]Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA

*lromo@broadinstitute.org

Noncoding variants are increasingly recognized as contributors to rare disease, yet their clinical interpretation remains challenging due to limited annotation of regulatory elements. I previously showed that 3′UTR variants overlapping microRNA sites, RNA binding protein motifs, and polyadenylation signals are enriched for pathogenicity in ClinVar. To extend these findings across other noncoding regions and translate them into improved diagnosis, we annotated noncoding variants from multiple sources: the GREGoR rare disease cohort, pathogenic and benign variants from ClinVar, *de novo* variants from autism cases and controls in denovo-db, and common functional variants from eQTL and GWAS studies. Variants from GREGoR and denovo-db that overlapped regulatory elements showed higher conservation, constraint, and pathogenicity scores than those outside these regions, indicating enrichment for disease relevance. Both pathogenic and functional variants were more likely than benign or nonfunctional variants to fall within regulatory elements. We next sought to distinguish common functional variants from rare pathogenic 5′UTR variants. Known pathogenic ClinVar variants were about 10 times more likely than benign variants to occur in uORF start or stop codons, whereas common functional variants had up to 14-fold higher odds of being located in promoters, CpG islands, or open chromatin. Pathogenic and functional variants were largely in distinct genes: pathogenic variants were more frequent in OMIM genes, whereas GWAS variants were most likely to occur in constrained genes. *De novo* variants in children with autism were up to 2.8 times more likely than those in controls to fall within promoters, enhancers, or transcription factor motifs, particularly in conserved and constrained regions. Together, these results indicate that although both pathogenic and functional variants cluster within regulatory elements, variants that alter uORFs in OMIM genes are most likely to be associated with rare disease. We applied these insights to prioritize noncoding variants from the GREGoR cohort, identifying two novel pathogenic variants to date. A *CAPN3* 3′UTR variant found in trans with a coding pathogenic allele segregated with limb-girdle muscular dystrophy in three siblings and was predicted to strengthen RBFOX binding, a protein linked to neuromuscular disease. Functional studies validated the variant's pathogenicity. We also identified a *MEF2C* 5′UTR variant in a proband with a neurodevelopmental disorder that creates a uORF with a premature stop codon and strong Kozak sequence, consistent with disrupted translation. These findings demonstrate the value of systematic regulatory annotation for identifying pathogenic noncoding variants missed by conventional analyses and underscore the need for improved computational tools to interpret regulatory variation in rare disease.

Beyond splice site identification: machine learning models for prediction of variant effects in splicing quantification

Matthew Runyan[1, *], Saumya Gupta[1], Jennifer Dy[2], Predrag Radivojac[3], Peter Castaldi[4], and Ayan Paul[1, 4,]

[1]The Institute for Experiential AI, Northeastern University, Boston, MA, USA

[2]Department of Electrical and Computer Engineering, Northeastern University, Boston, MA, USA

[3]Khoury College of Computer Sciences, Northeastern University, Boston, MA, USA

[4]CDNM, Brigham & Women's Hospital, Harvard Medical School, Boston, MA, USA

*runyan.m@northeastern.edu

Abstract

Predictions of the effects of genetic variants on pre-mRNA splicing are essential for clinical diagnostics, personalized medicine, and deepening our understanding of the underlying splicing mechanisms. Although computational models such as SpliceAI, Pangolin, and SpliceTransformer have achieved high accuracy in identifying splice sites, their predictive performance decreases when tasked with quantifying splice site usage (SSU) against RNA sequencing data.

In this study, we used an in-house dataset comprising 182 human airway epithelial cell (HAEC) samples, each with paired RNA sequencing and genotyping data, to evaluate the extent to which splicing models generalize to an unseen tissue and across individuals. Using personal transcriptomic sequences, we assessed model predictions against experimentally observed SSU values. We then trained our own Dilated Convolutional Neural Networks (DCNNs) on this HAEC dataset, specifically comparing a model trained on personal transcriptomic sequences to a model trained with the same SSU targets but using only sequences extracted from the reference genome.

On 79 held-out individuals and chromosomes, HAEC-trained models achieved a coefficient of determination (R^2) of 84.2% compared 75.7% and 71.4% for SpliceAI and Pangolin respectively, demonstrating poor cross-tissue generalization for SSU prediction. We further evaluated SpliceAI and Pangolin against our own models using splicing quantitative trait loci (sQTL) analysis in the HAEC dataset and validated variant effect predictions with data from massively parallel splicing assays (MPSA). Using data from the Vex-Seq MPSA we evaluated the ability of SpliceAI, Pangolin, and our HAEC-trained models to predict the percent spliced in (PSI) of cassette exons, both in the presence and absence of specific variants. From these predictions we computed the change in predicted PSI due to each variant and correlated this predicted ΔPSI to the experimentally observed ΔPSI. The HAEC-trained model using personal transcriptomic sequences achieved the highest Pearson correlation coefficient of 0.58, while the same model trained on reference transcriptome sequences reached 0.57. In comparison, both SpliceAI and Pangolin achieved a correlation of 0.56.

In summary, training with personal transcriptomes led to only modest gains over reference-based training. These findings support the need for tissue-specific training and model architectures designed to learn variant effects from personal transcriptomes.

Predicting interaction-specific protein–protein interaction perturbations by missense variants with MutPred-PPI

Ross Stewart[1], Florent Laval[2,3,4,5,6], Michael A. Calderwood[2,3,4], Matthew Mort[7], David N. Cooper[7], Marc Vidal[2,3], Predrag Radivojac[1,*]

[1]Khoury College of Computer Sciences, Northeastern University, Boston, MA, USA
[2]Center for Cancer Systems Biology (CCSB), Dana-Farber Cancer Institute, Boston, MA, USA
[3]Department of Genetics, Blavatnik Institute, Harvard Medical School, Boston, MA, USA
[4]Department of Cancer Biology, Dana-Farber Cancer Institute, Boston, MA, USA
[5]TERRA Teaching and Research Centre, University of Liège, Gembloux, Belgium
[6]Laboratory of Viral Interactomes, GIGA Institute, University of Liège, Liège, Belgium
[7]Institute of Medical Genetics, School of Medicine, Cardiff University, Cardiff, UK

[*]Corresponding author: predrag@northeastern.edu

Abstract

Disruption of protein–protein interactions (PPIs) is a major mechanism of a variant's deleterious effect. While most variant effect predictors assess overall pathogenicity or protein properties such as stability, they rarely consider loss of specific interactions, particularly when the variant perturbs binding interfaces without significantly affecting protein stability. To address this problem, we present MutPred-PPI, a graph attention network that predicts interaction-specific (edgetic) effects of missense variants by operating on AlphaFold 3-based protein complex contact graphs with protein language model embeddings imposed upon nodes. We systematically evaluated our model with stringent group cross-validation as well as benchmark data recently collected within the IGVF Consortium. MutPred-PPI outperformed all baseline methods across all evaluation criteria, achieving an AUC of 0.85 on seen proteins and 0.72 on previously unseen proteins in cross-validation, demonstrating strong generalizability despite scarce training data. To demonstrate biomedical relevance, we applied MutPred-PPI to variants from ClinVar, HGMD, COSMIC, gnomAD, and two *de novo* neurodevelopmental disorder-linked datasets. Disease-associated variants from ClinVar and HGMD showed strong enrichment for both quasi-null and edgetic effects, whereas population variants from gnomAD increasingly preserved interactions with higher allele frequencies. Notably, we observed a strong edgetic disruption signature in highly recurrent cancer variants from both the full COSMIC dataset and a subset of variants from oncogenes. Recurrent tumor suppressor gene variants and autism spectrum disorder-associated variants exhibited moderate quasi-null enrichment, whilst neurodevelopmental disorder-linked variants showed a weak edgetic disruption signature. These results indicate distinct PPI perturbation mechanisms across disease types and show that MutPred-PPI captures functionally relevant molecular effects of pathogenic variants.

Integrating Structural, Coevolutionary, and Protein Language Models for CAGI7 Challenge Predictions

Matsvei Tsishyn, Hugo Talibart, Pauline Hermans, Gabriel Cia,
Marianne Rooman and Fabrizio Pucci*

Computational Biology and Bioinformatics, Université Libre de Bruxelles,
1050 Bruxelles, Belgium

*email : Fabrizio.Pucci@ulb.be

In our participation in CAGI7, we developed several strategies for the different challenges in which we took part. Our key approach consists of integrating various computational methods, which are combined depending on the challenge being. We employed three main types of features for prediction: physics-based statistical potentials, (co)evolutionary methods, and protein language models.

Statistical potentials [1], developed in our laboratory over more than a decade, are mean-force potentials derived from datasets of well-resolved protein 3D structures using the inverse Boltzmann law. These potentials allow us to compute the folding free energy of proteins. The (co)evolutionary approaches used are: RSALOR [2] that is a method we develop to estimate mutation fitness from multiple sequence alignments and is based on the difference in frequency between wild-type and mutant residues, weighted by the solvent accessibility of the residue; Structured-DCA [3] a new structure-informed pseudolikelihood maximization direct coupling analysis that we are developing that reduces the number of residue couplings in the DCA model by incorporating 3D structural information. Finally, we employed SaProt [4], a protein language model that also considers the 3D structure of the input protein.

All these different features were integrated using two main strategies: a simple linear combination and a shallow neural network. Models were trained on datasets specifically curated for each challenge. We first searched the literature for deep mutational scanning (DMS) data derived from assays like those used in the challenges. In parallel, we employed a curated subset of DMS data from ProteinGym [5], selected according to the properties that needed to be predicted in each challenge.

References

[1] Dehouck, Yves, Dimitri Gilis, and Marianne Rooman. "A new generation of statistical potentials for proteins." *Biophysical journal* 90.11 (2006): 4010-4017.

[2] Tsishyn, Matsvei, et al. "Residue conservation and solvent accessibility are (almost) all you need for predicting mutational effects in proteins." *Bioinformatics* 41.6 (2025): btaf322.

[3] Tsishyn, Matsvei et al., Structure-informed Direct Coupling Analysis to predict mutational landscape, *in preparation*

[4] Su, Jin, et al. "Saprot: Protein language modeling with structure-aware vocabulary." *BioRxiv* (2023): 2023-10.

[5] Notin, Pascal, et al. "Proteingym: Large-scale benchmarks for protein fitness prediction and design." *Advances in Neural Information Processing Systems* 36 (2023): 64331-64379.