Evaluation of performance of leading algorithms for variant pathogenicity predictions and designing a combinatory predictor method: Application to Rett syndrome variants

Satishkumar Ganakammal, Emil Alexov

1 School of Nursing, Clemson University, Clemson, SC, United States
2 Department of Physics, Clemson University, Clemson, SC, United States

Corresponding Author: Emil Alexov
Email address: ealexov@clemson.edu

Background Genomics diagnostic tests are done for a wide spectrum of complex genetics conditions such as autism and cancer. The growth of technology has not only aided in successfully decoding the genetic variants that causes or trigger these disorders. However, interpretation of these variants is not a trivial task even at a level of distinguish pathogenic versus benign variants. Methods: We used the clinically significant variants from ClinVar database to evaluate the performance of fourteen most popular in-silico predictors using supervised learning methods. We implemented a feature selection and random forest classification algorithm to identify the best combination of predictors to evaluate the pathogenicity of a variant. Finally, we have also utilized this combination of predictors to reclassify the variants of unknown significance in MeCP2 gene that are associated with the Rett syndrome. Results: The results from analysis shows an optimized selection of prediction algorithm and developed a combinatory predictor method. Our combinatory approach of using both best performing independent and ensemble predictors reduces any algorithm biases in variant characterization. The reclassification of variants (such as VUS) in MECP2 gene associated with RETT syndrome suggest that the combinatory in-silico predictor approach had a higher success rate in categorizing their pathogenicity.
Evaluation of performance of leading algorithms for variant pathogenicity predictions and designing a combinatory predictor method: Application to Rett syndrome variants

Satishkumar RanganathanGanakammal¹, Emil Alexov¹,²

¹ Department of Healthcare Genetics, School of Nursing, Clemson University, Clemson, SC 29634, USA
² Department of Physics and Astronomy, Clemson University, Clemson, SC 29634, USA

Corresponding Author:
Emil Alexov
Email address: ealexov@clemson.edu

Abstract

Background
Genomics diagnostic tests are done for a wide spectrum of complex genetics conditions such as autism and cancer. The growth of technology has not only aided in successfully decoding the genetic variants that causes or trigger these disorders. However, interpretation of these variants is not a trivial task even at a level of distinguish pathogenic versus benign variants.

Methods:
We used the clinically significant variants from ClinVar database to evaluate the performance of fourteen most popular in-silico predictors using supervised learning methods. We implemented a feature selection and random forest classification algorithm to identify the best combination of predictors to evaluate the pathogenicity of a variant. Finally, we have also utilized this combination of predictors to reclassify the variants of unknown significance in MeCP2 gene that are associated with the Rett syndrome.

Results:
The results from analysis shows an optimized selection of prediction algorithm and developed a combinatory predictor method. Our combinatory approach of using both best performing independent and ensemble predictors reduces any algorithm biases in variant characterization. The reclassification of variants (such as VUS) in MECP2 gene associated with RETT syndrome suggest that the combinatory in-silico predictor approach had a higher success rate in categorizing their pathogenicity.

Introduction
With advances in genomic sequencing, molecular genomics has quickly become a standard in clinical genetics and diagnostics. Molecular genetic testing involves the identification of variants in clinical actionable regions of a single gene or multiple genes that can cause a genetic disorder. Gene based molecular diagnostics tests have the ability to contribute to a wide range of testing types such as screening for specific conditions such as autism or cancer. As the degree of severity of the disease depends on the genes affected and its associated variants along with their phenotype characteristics, multi-gene screening provides a gateway to analyze a set of genes that are found to be associated with a specific or multiple phenotype all at once (Niroula & Vihinen, 2017). The field of genomics has provided a continuous evolving platform to decode the human genome.

Advances in molecular genomic technologies such as Next-generation sequencing that includes whole exome sequencing (WES) and whole genome sequencing (WGS) have aided in screening multiple genes in an instance to identify several single nucleotide variants (SNVs) that contribute to a disease (Health & Alliance, 2010). Many disease phenotypes have been linked to the missense variants. They are considered as most clinically relevant as they alter the amino acid encoding a protein that can affect the gene function. These SNVs are not only involved in disease causing but also play an important role in altering biological processes such as transcriptional regulation, splicing (Thusberg & Vihinen, 2009).

The SNVs identified from methods like WGS or WES are evaluated based on the metadata obtained from variant annotation process that is a part of the bioinformatics workflow using information from various data sources such as Online Mendelian Inheritance in Man (OMIM) (Hamosh et al., 2005) and Human Gene Mutation Database (HGMD) (Stenson et al., 2014) databases. Using this annotated information, the variants are classified into pathogenic, benign, likely pathogenic or benign, variant of uncertain significance (VUS) and incidental findings (Ifs) based on American College of Medical Genetics and Genomics (ACMG) recommendations (Richards et al., 2015). The process of characterizing a particular variant’s clinical relevance such as pathogenic (disease causing) or non-pathogenic (non-disease causing or benign) poses a challenge due to issues such as differences in information from the bioinformatics workflows, limited availability of computational resources and the lack of trained professionals, despite that various computational algorithms have been developed to predict the clinical pathogenicity of variant based features such as homology, conservation based on evolution, protein function etc. (Dong et al., 2014).

In this study we employ machine learning strategies on variants from ClinVar database (Landrum et al., 2015) to evaluate and identify the best combination of in-silico prediction algorithms to implement a best performing combinatory predictor method to characterize its pathogenicity. The results can provide a framework for bioinformaticists and molecular genomicists to review the clinical relevance of a variant by minimizing both false positive and false negative predictions.
This also provides a benchmark set of predictors that could be used to determine and reclassify the variants of unknown clinical significance.

**Overview of In-silico Prediction Algorithms**

Most of the computational methods use prediction features and then identify and implement the best performing algorithm on a training data set to classify and curate the pathogenicity of variant. (Vihinen, 2012) Some predictors aim at predicting if the variant is disease-causing, others focus on predicting molecular effects caused by the mutation (Kucukkal et al., 2015; Petukh, Kucukkal & Alexov, 2015; Peng & Alexov, 2016; Peng, Alexov & Basu, 2019). ACMG has drafted policy statement and guidelines for categorizing variants according to which a variant should have multiple computational evidence to support its deleterious effect from these computational algorithms or predictors. There are two types of in-silico predictors (Table 1), (i) Independent predictors (SIFT (Sorting Intolerant From Tolerant) (Ng & Henikoff, 2003), Polyphen2 (Polymorphism Phenotyping V-2) (Adzhubei et al., 2010), PROVEAN (Protein Variation Effect Analyzer) (Choi et al., 2012), LTR (Likelihood ratio test) (Chun & Fay, 2009), Mutation Taster (Schwarz et al., 2010), Mutation Assessor (Reva, Antipin & Sander, 2011), FATHMM (Shihab et al., 2014), VEST3 (Carter et al., 2013)) employs computational algorithm that considers unique features to determine the deleteriousness of a variant (ii) Ensemble predictor (REVEL (Ioannidis et al., 2016), M-CAP (Mendelian Clinically Applicable Pathogenicity) (Jagadeesh et al., 2016), MetaLR (Dong et al., 2014), MetaSVM (Dong et al., 2014), CADD (Combined Annotation Dependent Depletion) (Rentzsch et al., 2018), Eigen (Ionita-Laza et al., 2016)) includes computation algorithms that uses collected features from various independent predictors to determine the pathogenicity of a variant. These prediction methods use different computational algorithms (Markov model, evolutionary conservation, random forest etc.) so in order to eliminate the algorithm biases it is advisable to use multiple prediction algorithms for variant evaluation.(Richards et al., 2015)

**Materials & Methods**

**Dataset:**

The ClinVar variant data is downloaded in a tab separated format from the ClinVar database (a public archive with interpretations of clinically relevant variants) FTP site (ftp://ftp.ncbi.nlm.nih.gov/pub/clinvar/) (Landrum et al., 2015). A customized perl script was written to parse single nucleotide variants corresponding to the “GRCh37” build. Table 2 shows the counts variants from the cleaned data categorized with clinical relevance such as “pathogenic”, “benign”, “likely pathogenic”, “likely benign” and variant of uncertain significance (VUS). For this paper, we choose only “pathogenic” and benign” and then apply purging procedure to eliminate cases where there is no strong evidence to be classified as “pathogenic” or “benign”. The purging was done by reading number of submitters (NOS) and submitter categories (SC) and then applying the following filters: (a) for pathogenic mutations taken from ClinVar we require that NOS > 2 & SC = 3; (b) for non-pathogenic/benign mutation the filter was NOS > 3 & SC >= 2
This reduced the number of pathogenic mutations from 36536 to 2123, and benign mutations from 7249 to 2231. This reduced dataset is termed “golden set”. The purging had two outcomes: reduced the noise of potentially wrong classifications and the number of pathogenic and benign mutations became very similar. This dataset serves for two purposes (i) as cross validation data set to test the performance of all the 14 in-silico algorithms evaluated (ii) and as a training data set to aid in re-classification of the uncategorized variants types such as VUS or conflicting interpretation variants.

Another dataset, called Rett syndrome dataset, consists of variants in MeCP2 gene that are associated with Rett syndrome (Zoghbi, 2016; Gold et al., 2017). The data was extracted from downloaded ClinVar database. This data set has all the differently clinically categorized variants that includes 64 pathogenic and 1 benign variant and 115 variants classified as like benign/pathogenic, VUS and conflicting interpretations. Out of the 115 variants we were able to clean up our testing to 101 variants that has associated in-silico predictor score available. They will be subjected to our best set of in-silico predictors and reclassified as pathogenic or benign.

**Feature Extraction:**

We annotated all the variants from our prepared dataset using the dbNSFP data source v2.9.3 ([https://sites.google.com/site/jpopgen/dbNSFP](https://sites.google.com/site/jpopgen/dbNSFP)) (Liu, Jian & Boerwinkle, 2011). This data source includes scores from all the in-silico predictors along with allele frequency information from various population databases. A customized perl script is used to extract the consolidated in-silico scores from both dependent and independent predictors such as SIFT, Polyphen2, LTR, Mutation Taster, Mutation Assessor, FATHMM, PROVEAN, VEST3, MetaSVM, MetaLR, M-CAP, Revel, CADD, Eigen. These scores are used as features for our features ranking and performance evaluation algorithms.

**Features ranking and performance evaluation:**

We used the scores from the 14 in-silico predictors as features to access their performance. We evaluated the in-silico predictors on the ability to distinguishing the variants of our dataset into pathogenic or benign based on the statistics collected from the confusion matrix. We used Weka (v3.8.2) (Hall et al., 2009) to collect statistics about accuracy, sensitivity, specificity, precision, F-measure and MCC (Mathews correlation coefficient) calculated using the number of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) (**Table 4**).

**Identification of best in-silico predictor set:**

For the development of the best predictor set, firstly we evaluated the two class of attributes or features (independent and ensemble predictors scores) separately using the ranker attribute evaluation method. This provides with a list of the best performing in-silico algorithms. Secondly, we identified the best combination of in-silico predictors that includes top performing independent
and ensemble predictors that can best classify the variant as either benign or pathogenic. The
“ranker” option nested under the classifier attribute evaluation method is used for ranking features
(in-silico scores), it is a fast and precise method that considers only relevant attributes and
eliminates both irrelevant and redundant features that ranking our methods based on more on
correlation. Thus, the algorithm ranks the features based on their strength of classification. We
also use Random forest as a classifier method along with the Ranker evaluator method to rank and
evaluate the in-silico predictor based on performance.

Results
Selection of best classification algorithms:
The predictors’ scores on the golden dataset with 4,354 variants bearing stronger evidence to be
categorized as pathogenic and benign were used as an input for the machine learning algorithms.
For the first step of our analysis we applied the best performing classification methods for the
evaluation of our dataset. Table 5 shows the statistics. Based on our findings, we identified
Random forest method as the best classifier method when compared to others such as Naïve Bayes,
Classification via Regression, LibSVM with 97% accuracy.

Evaluation of the performance of in-silico prediction methods:
After identifying Random forest as the best classifier method, we evaluated the performance of the
in-silico predictors separately based on the strength of classification of a variant into benign or
pathogenic class. Comparing the statistics obtained from the classification method (Figure 1) we
identified that ensemble or the dependent predictor out performed almost all the independent
predictor algorithms with higher sensitivity and accuracy. VEST3 is the only independent predictor
that has seems to have a sensitivity and accuracy comparable to the ensemble predictors. Even the
comparison based on MCC values that is used for evaluation of imbalanced data displays the same
trend where the ensemble predictors fared better than the independent predictors.

We also performed evaluation of dependent predictors and independent/ensemble predictors
separately using the 10-fold cross validation strategy by implementing random forest method to
check for the best performing classification predictors based on accuracy and other statistics of
predictions. The ensemble predictors outperform the independent predictors with an accuracy of
~97 % along with higher sensitivity, specificity and MCC values (Figure 2).

Identification of the best performing in-silico predictor set
The results from the feature evaluation of the ranking attribute methods for both the independent
and ensemble predictors separately identified the best in-silico predictors that can now be used to
classify the variant data better into pathogenic and benign sub-categories.

VEST3, LTR, Polyphene2 and PROVEAN are identified as the top four ranked independent in-
silico predictors and CADD, Eigen, MetaSVM and REVEL are identified as the top four ranked
ensemble in-silico predictors. **Figure 1** shows that these in-silico predictors did show higher sensitivity, MCC and accuracy trends too compared to the other in-silico predictors in their respective category.

The classification of SNVs with the top four predictors from both independent and ensemble predictor categories together shows a better and stronger evidence to evaluate variant pathogenicity. The higher rate of observed accuracy, sensitivity and the MCC statistics from the classification of our data with just the selected 8 features proves that our combined in-silico predictor set can be highly reliable with comparatively minimal algorithmic biases (**Table 6**). **Figure 3** shows that our identified combination of the predictors outperforms previously proposed combination proposed by Li et al., 2018 that includes just two predictors (REVEL & VEST3).

**Rett syndrome variants**

The 197 variants associated with Rett syndrome collected from the ClinVar database includes 101 variants which are categorized into likely benign/pathogenic, uncertain significant (VUS) and conflicting interpretations of pathogenicity classes, along with 64 pathogenic and 1 benign variant.

We used our best in-silico predictor set to reclassify the above (64 pathogenic and 1 benign) variants either as pathogenic or benign with an average classification accuracy of 89% and 100%, respectively (**Table 7**). This assures the performance of the proposed set of in-silico predictors. Furthermore, **Table 7** summarizes the fringe variants re-classified, with our best set of in-silico predictors out of which 60% was classified into pathogenic category and 40% was classified into benign category.

**Discussion**

The advances in computer algorithm had been widely utilized in the evaluation of the pathogenicity of a variant. We evaluated the performance of 14 prominent in-silico computational algorithm methods with 4354 SNVs from purged ClinVar database (golden dataset). We also evaluated the performance of 8 independent and 6 ensemble predictors that led us to identify the best combination of in-silico predictors that can categorize variants into either disease causing or not. Our initial investigation revealed that the ensemble prediction algorithms outperformed the independent algorithm with a higher accuracy of variant categorization.

Our individual assessment of the in-silico prediction methods shows that VEST3 is the best performing independent predictor method with high accuracy of classifying variants. However, the main limitation is that this algorithm is based on prioritization of missense variants thus creating a partisanship biases in evaluating the non-missense variants. Similarly, Eigen and CADD are the best performing empirical algorithms based on accuracy of classification which is also highly influenced by the algorithm constrain that decreases the sensitivity and specificity of variant characterization. The pathogenicity of variant can also be associated with the different scoring strategies used by either supervised or unsupervised learning methods. This provides us with a
strong platform to investigate a combinatory approach that includes both dependent and empirical predictors to evaluate the pathogenicity of a variant.

Although there are many studies that performed comparative investigation to identify the best performing in-silico predictor method, (Li et al., 2018) in their study displayed the effectiveness of a combinatory approach where the combination of two in-silico predictors, VEST3 and REVEL displayed better overall performance in characterization of clinically significant missense variants. This combination when extended to both missense and non-missense variants displayed less accuracy, sensitivity and specificity compared to the just the empirical predictors. The results from feature selection analysis identified the best combination of independent and empirical predictors that can distinguish and characterize the variant pathogenicity.

To select the best combination of predictors, we first used the attribute ranking method for overall ranking of the predictors, followed by combination predictors using each in-silico predictor sub-group (independent and ensemble). We picked the combination with the least number of predictors with the highest prediction accuracy. Based on the statistics obtained (a few combinations are highlighted in supplementary Table 1) the combination of top 4 methods had the highest accuracy, sensitivity and MCC values. Thus we combined the top four performing in-silico predictors from both the empirical and the conventional methods (independent methods: VEST3, LTR, Polyphene2 PROVEAN; ensemble methods: CADD, Eigen, MetaSVM and REVEL) that yielded an accuracy of 97% which is similar to the accuracy yielded by just the empirical predictors, while providing information on a variant with a minimized biased evaluation. It even outperformed the ReVe (Revel and Vest3) combination from Li et al., 2018 with higher sensitivity and accuracy measures (Figure 3) (Supplementary Table 1). During our selection we noticed that the empirical predictors MetaLR and REVEL both exhibited similar accuracy, sensitivity and specificity but we included REVEL into our set as it has been validated with larger sample set and has exhibited greater performance in classifying missense variants.

After we demonstrated that the algorithm performs well, it was applied to reclassify the variants associated with RETT syndrome listed as uncertain or conflicting clinical significance. This reclassified set can be used to guide further clinical investigation for mutants linked with Rett syndrome along with studies about the effects of mutation on wild type characteristics of the corresponding protein.

**Conclusions**

In summary, our combinatory approach of using both best performing independent and empirical predictors reduces any algorithm biases in variant characterization. Our robust training dataset composed of ClinVar variants filtered based on strong evidences for pathogenic and benign characteristics can reduce the false positive and true negative results. Also, similar filtering approaches for data preparation can be used in development of new methods for assessing the functional effect of a variant. Though in-silico predictors are just one of data point in evaluation
of variant pathogenicity along with other information such as allele frequency, our predictor set
will aid in consolidated balanced prediction thus increases the confidence of evaluation. This also
could provide sufficient evidence for clinical genomicist and researchers to understand and
evaluate the pathogenicity of variants whose clinical relevance is unknown.

References

1. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov
   AS, Sunyaev SR. 2010. A method and server for predicting damaging missense mutations.
   Nature methods 7:248.
2. Carter H, Douville C, Stenson PD, Cooper DN, Karchin R. 2013. Identifying Mendelian
disease genes with the variant effect scoring tool. BMC genomics 14:S3.
3. Choi Y, Sims GE, Murphy S, Miller JR, Chan AP. 2012. Predicting the functional effect
of amino acid substitutions and indels. PloS one 7:e46688.
4. Chun S, Fay JC. 2009. Identification of deleterious mutations within three human genomes.
   Genome research 19:1553–1561.
5. Dong C, Wei P, Jian X, Gibbs R, Boerwinkle E, Wang K, Liu X. 2014. Comparison and
integration of deleteriousness prediction methods for nonsynonymous SNVs in whole
exome sequencing studies. Human molecular genetics 24:2125–2137.
6. Gold WA, Krishnarajy R, Ellaway C, Christodoulou J. 2017. Rett syndrome: a genetic
update and clinical review focusing on comorbidities. ACS chemical neuroscience 9:167–
176.
7. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. 2009. The WEKA
data mining software: an update. ACM SIGKDD explorations newsletter 11:10–18.
8. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. 2005. Online Mendelian
Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders.
Nucleic acids research 33:D514–D517.
9. Health D of CD of, Alliance G. 2010. Understanding genetics: A District of Columbia
guide for patients and health professionals. Genetic Alliance.
10. Ioannidis NM, Rothstein JH, Pejaver V, Middha S, McDonnell SK, Baheti S, Musolf A,
   Li Q, Holzinger E, Karyadi D. 2016. REVEL: an ensemble method for predicting the
   pathogenicity of rare missense variants. The American Journal of Human Genetics 99:877–
   885.
11. Ionita-Laza I, McCallum K, Xu B, Buxbaum JD. 2016. A spectral approach integrating
functional genomic annotations for coding and noncoding variants. Nature genetics
   48:214.
12. Jagadeesh KA, Wenger AM, Berger MJ, Guturu H, Stenson PD, Cooper DN, Bernstein
   JA, Bejerano G. 2016. M-CAP eliminates a majority of variants of uncertain significance
   in clinical exomes at high sensitivity. Nature genetics 48:1581.
13. Kucukkal TG, Petukh M, Li L, Alexov E. 2015. Structural and physico-chemical effects of
disease and non-disease nsSNPs on proteins. Current opinion in structural biology 32:18–
24.
14. Landrum MJ, Lee JM, Benson M, Brown G, Chao C, Chitipiralla S, Gu B, Hart J, Hoffman
   D, Hoover J. 2015. ClinVar: public archive of interpretations of clinically relevant variants.
   Nucleic acids research 44:D862–D868.
15. Li J, Zhao T, Zhang Y, Zhang K, Shi L, Chen Y, Wang X, Sun Z. 2018. Performance
evaluation of pathogenicity-computation methods for missense variants. Nucleic acids
16. Liu X, Jian X, Boerwinkle E. 2011. dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. *Human mutation* 32:894–899.

17. Ng PC, Henikoff S. 2003. SIFT: Predicting amino acid changes that affect protein function. *Nucleic acids research* 31:3812–3814.

18. Niroula A, Vihinen M. 2017. Predicting severity of disease-causing variants. *Human mutation* 38:357–364.

19. Peng Y, Alexov E. 2016. Investigating the linkage between disease-causing amino acid variants and their effect on protein stability and binding. *Proteins: Structure, Function, and Bioinformatics* 84:232–239.

20. Peng Y, Alexov E, Basu S. 2019. Structural perspective on revealing and altering molecular functions of genetic variants linked with diseases. *International journal of molecular sciences* 20:548.

21. Petukh M, Kucukkal TG, Alexov E. 2015. On human disease-causing amino acid variants: statistical study of sequence and structural patterns. *Human mutation* 36:524–534.

22. Rehm HL, Berg JS, Brooks LD, Bustamante CD, Evans JP, Landrum MJ, Ledbetter DH, Maglott DR, Martin CL, Nussbaum RL. 2015. ClinGen—the clinical genome resource. *New England Journal of Medicine* 372:2235–2242.

23. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. 2018. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic acids research* 47:D886–D894.

24. Reva B, Antipin Y, Sander C. 2011. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic acids research* 39:e118–e118.

25. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, Grody WW, Hegde M, Lyon E, Spector E. 2015. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genetics in medicine* 17:405.

26. Schwarz JM, Rödelserger C, Schuelke M, Seelow D. 2010. MutationTaster evaluates disease-causing potential of sequence alterations. *Nature methods* 7:575.

27. Shihab HA, Gough J, Mort M, Cooper DN, Day INM, Gaunt TR. 2014. Ranking nonsynonymous single nucleotide polymorphisms based on disease concepts. *Human genomics* 8:11.

28. Stenson PD, Mort M, Ball E V, Shaw K, Phillips AD, Cooper DN. 2014. The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Human genomics* 133:1–9.

29. Thusberg J, Vihinen M. 2009. Pathogenic or not? And if so, then how? Studying the effects of missense mutations using bioinformatics methods. *Human mutation* 30:703–714.

30. Vihinen M. 2012. How to evaluate performance of prediction methods? Measures and their interpretation in variation effect analysis. In: *BMC genomics*. BioMed Central, S2.

31. Zoghbi HY. 2016. Rett syndrome and the ongoing legacy of close clinical observation. *Cell* 167:293–297.
Figure 1

Performance evaluations of fourteen in-silico predictors

The graphical representation of the major statistics obtained from the evaluation of all 14 in-silico predictors.
Figure 2

Performance comparison: independent vs empirical in-silico predictors

The graphical representation of the major statistics obtained from the evaluation of both independent (solid bars) and ensemble (grey bars) predictors.
Figure 3

Performance comparison: our approach vs ReVe

Comparison of the statistics obtained from the proposed combined set of independent (VEST3, LTR, Polyphen2 and PROVEAN) and ensemble or dependent (CADD, Eigen, MetaSVM and REVEL) predictors (solid bars) to the combination of REVEL and VEST as proposed by Li et al., 2018 (grey bars)
Manuscript to be reviewed

- Accuracy: Best set = 0.966, ReVe (Li et al.) = 0.808
- MCC: Best set = 0.933, ReVe (Li et al.) = 0.649
- Specificity: Best set = 0.95, ReVe (Li et al.) = 0.639
- Sensitivity: Best set = 0.982, ReVe (Li et al.) = 0.971
Table 1 (on next page)

Description of In-silico Predictors evaluated

Brief description of the fourteen in-silico predictors (both independent and empirical predictors) used in this study with pathogenicity cutoffs values.
**Table 1:** Description of In-silico Predictors evaluated.
A brief description of the fourteen in-silico predictors (both independent and empirical predictors) used in this study with pathogenicity cutoffs values.

| Predictor       | Description                                                                                                                                                                                                 | Pathogenicity cutoff |
|-----------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------|
| SIFT            | It uses MSA methodology that determines the probability that a missense variant is tolerated conditional on the most frequent amino acid being tolerated. (Ng & Henikoff, 2003)                                      | <0.049               |
| Polyphen2       | It calculates the normalized accessible surface area and changes in accessible surface propensity resulting from the amino acid substitution. (Adzhubei et al., 2010)                                      | >0.447               |
| LTR             | It uses heuristic methods to identify mutations that disrupt highly conserved amino acids within protein-coding sequences. (Chun & Fay, 2009)                                                                  | NA                   |
| Mutation Taster | It uses naive Bayes classifier to evaluate the pathogenicity of a variant based on information available from various databases. (Schwarz et al., 2010)                                                               | >0.5                 |
| Mutation Assessor| It uses the concept of evolutionary conservation that affects amino acid in protein homologs. (Reva, Antipin & Sander, 2011)                                                                                 | >1.935               |
| FATHMM          | It uses Hidden Markov Models (HMM) to assess the functionality of the candidate variant by incorporating a disease-specific weighting scheme. (Shihab et al., 2014),                                           | < -1.151             |
| PROVEAN         | It uses the concept of pairwise sequence alignment scores to predict the biological effect on the protein function. (Choi et al., 2012)                                                                     | < -.2.49             |
| VEST3           | It uses supervised learning method utilizing p-values generated by gene prioritization method to assess the functionality of mutations. (Carter et al., 2013)                                                    | NA                   |

**Empirical or Meta In-silico Predictors**

| Predictor  | Description                                                                                                           | Pathogenicity cutoff |
|------------|-----------------------------------------------------------------------------------------------------------------------|----------------------|
| MetaSVM    | It uses Support Vector Machine approach on the previous generated scores. (Dong et al., 2014)                              | >0                   |
| MetaLR     | It uses logistic regression model on the previous generated scores. (Dong et al., 2014)                                  | >0.5                 |
| M-CAP      | It uses gradient boosting trees method to analyze interactions between features to determine variant pathogenicity. (Jagadeesh et al., 2016) | NA                   |
| REVEL      | It combines all results from available prediction tools by using them as features to access the pathogenicity of a variant. (Ioannidis et al., 2016) | >0.75                |
| CADD       | It uses a c-score obtained by the integration of multiple variant annotation resources. (Rentzsch et al., 2018)          | >19                  |
| Eigen      | It uses a supervised approach to derive the aggregate functional score from various annotation resources. (Ionita-Laza et al., 2016) | NA                   |
**Table 2** (on next page)

Clinical relevance distribution of variants from ClinVar database

Counts of SNVs (Single Nucleotide Variants) from ClinVar Database (for build GRCh37) categorized based on major clinical relevance.
Table 2: Clinical relevance distribution of variants from ClinVar database.

Counts of SNVs (Single Nucleotide Variants) from ClinVar Database (for build GRCh37) categorized based on major clinical relevance.

| Clinical Relevance                  | Total Number of variants |
|------------------------------------|--------------------------|
| Pathogenic                         | 36536                    |
| Benign                             | 7249                     |
| Likely Pathogenic                  | 2105                     |
| Likely Benign                      | 17295                    |
| Variant of unknown Significance (VUS) | 135534                  |
**Table 3** (on next page)

Proposed Golden dataset set

The golden data set that includes pathogenic and benign variants obtained by filtering the ClinVar SNVs (build GRCh37) based on the number of submitters (NOS) and submitter categories (SC).
Table 3: Proposed Golden dataset set.
The golden data set that includes pathogenic and benign variants obtained by filtering the ClinVar SNVs (build GRCh37) based on the number of submitters (NOS) and submitter categories (SC).

| Clinical Relevance | Total Number of variants | Criteria               |
|--------------------|--------------------------|------------------------|
| Pathogenic         | 2123                     | NOS > 2 & SC = 3       |
| Benign             | 2231                     | NOS > 3 & SC >= 2      |
| Total              | 4354                     |                        |
Table 4 (on next page)

Statistical measure from our supervised learning method

Various statistics values calculated from our performance evaluation and classification analysis from Weka Software.
Table 4: 
Statistical measure from our supervised learning method.
Various statistics values calculated from our performance evaluation and classification analysis from Weka Software.

| Statistics   | Formula                                                                 |
|--------------|--------------------------------------------------------------------------|
| Sensitivity  | $\frac{TP}{(TP + FN)}$                                                  |
| Specificity  | $\frac{TN}{(TN + FP)}$                                                  |
| Precision    | $\frac{TP}{(TP + FP)}$                                                  |
| F -measure   | $\frac{2 \times \text{Precision} \times \text{recall}}{(\text{Precision} + \text{recall})}$ |
| MCC          | $\frac{TP \times TN - FN \times FP}{\sqrt{(TP + FN)(TP + FP)(TN + FN)(TN + FP)}}$ |
Table 5 (on next page)

Summary of various supervised learning method

Statistics calculated on our cross-validation dataset by applying different machine learning algorithms to identify the best methods for feature evaluation.
Table 5: Summary of various supervised learning method.
Statistics calculated on our cross-validation dataset by applying different machine learning algorithms to identify the best methods for feature evaluation.

| Classification Algorithm | Sensitivity | Specificity | Precision | Recall  | F-Measure | MCC   | Accuracy |
|--------------------------|-------------|-------------|-----------|---------|-----------|-------|----------|
| Random Forest            | 0.985       | 0.952       | 0.956     | 0.985   | 0.970     | 0.938 | 0.969    |
| Naive Bayes              | 0.905       | 0.911       | 0.914     | 0.905   | 0.909     | 0.815 | 0.907    |
| Classification via Regression | 0.957 | 0.944       | 0.948     | 0.957   | 0.953     | 0.902 | 0.951    |
| LibSVM                   | 0.940       | 0.930       | 0.934     | 0.940   | 0.937     | 0.870 | 0.953    |
Summary statistics of our combinatory approach

Statistics obtained by applying our classifier to the golden dataset with proposed combined set of independent (VEST3, LTR, Polyphen2 and PROVEAN) and ensemble or dependent (CADD, Eigen, MetaSVM and REVEL) predictors.
Table 6: Summary statistics of our combinatory approach.

Statistics obtained by applying our classifier to the golden dataset with proposed combined set of independent (VEST3, LTR, Polyphene2 and PROVEAN) and ensemble or dependent (CADD, Eigen, MetaSVM and REVEL) predictors.

| Predictors | Classification Algorithm | Sensitivity | Specificity | Precision | Recall | F-Measure | MCC | Accuracy |
|------------|--------------------------|-------------|-------------|-----------|--------|-----------|-----|----------|
| VEST3, LTR, Polyphene2, PROVEAN, CADD, Eigen, MetaSVM and REVEL | Random Forest | 0.982 | 0.950 | 0.954 | 0.982 | 0.968 | 0.933 | 0.966 |
Table 7 (on next page)

Reclassification of the MECP2 variants

Variants that was previously classified as likely benign/pathogenic, uncertain significant (VUS) and conflicting interpretations of pathogenicity classes was reclassified using our golden dataset (as training dataset) along with benchmarking against “pathogenic” and “benign” mutations.
Table 7: Reclassification of the MECP2 variants

Variants that was previously classified as likely benign/pathogenic, uncertain significant (VUS) and conflicting interpretations of pathogenicity classes was reclassified using our golden dataset (as training dataset) along with benchmarking against “pathogenic” and “benign” mutations.

| Clinical Significance     | Total Variants | Classification on best in-silico predictors | Success rate |
|----------------------------|----------------|---------------------------------------------|--------------|
|                            |                | Benign | Pathogenic |                  |              |
| Pathogenic                 | 64             | 7      | 57         | 89%              |
| Benign                     | 1              | 1      | 0          | 100%             |
| Likely Benign              | 10             | 9      | 1          | NA               |
| Likely Pathogenic          | 11             | 2      | 9          | NA               |
| Uncertain Significance     | 69             | 25     | 44         | NA               |
| Conflicting Interpretation | 11             | 5      | 6          | NA               |