Network-based pathogenicity prediction for variants of uncertain significance

Authors: Mayumi Kamada[1†], Atsuko Takagi[1†], Ryosuke Kojima[1†], Yoshihisa Tanaka[2,5], Masahiko Nakatsui[3], Noriko Tanabe[4], Makoto Hirata[4], Teruhiko Yoshida[4], Yasushi Okuno[1,5]*

Affiliations:

[1] Graduate School of Medicine, Kyoto University; Kyoto, 606-8507, Japan.
[2] Graduate School of Pharmaceutical Sciences, Kyoto University; Kyoto 606-8501, Japan
[3] Yamaguchi University Graduate School of Medicine, Yamaguchi University Hospital; Yamaguchi, 753-8511, Japan
[4] National Cancer Center Hospital; Tokyo, 104-0045 Japan.
[5] HPC- and AI-driven Drug Development Platform Division, RIKEN Center for Computational Science; Hyogo 650-0047, Japan

† These authors contributed equally to this work.
* Corresponding author. Email: okuno.yasushi.4c@kyoto-u.ac.jp

Abstract:

While the number of genome sequences continues to increase, the functions of many detected gene variants remain to be identified. These variants of uncertain significance constitute a major barrier to precision medicine. Although many computational methods have been developed to predict the function of these variants, they all rely on individual gene features and do not consider complex molecular relationships. Here we develop PathoGN, a molecular network-based approach for predicting variant pathogenicity. PathoGN significantly outperforms existing methods using benchmark datasets. Moreover, PathoGN successfully predicts the pathogenicity of 3,994 variants of uncertain significance in the real-world database ClinVar and designates potential pathogenicity. This is the first computational method for the clinical interpretation of variants using biomolecular networks, and we anticipate our method to be broadly useful for the clinical interpretation of variants and for assigning biological function to unknown variants at the genomic scale.

Main Text:

As genome sequencing accelerates, the functional interpretation of biological phenomena hidden in the genomic code is becoming extremely important. Recently, genomic medicine has begun to be integrated in a clinical setting with genomic information used to optimize medical care to the individual patient. Diagnostic and treatment decisions can be made based on patient genome analysis, literature, and database research on disease-related genes, alongside family history information and the patient's clinical background. This process is referred to as the "clinical interpretation" of genomic data. Since the information and criteria used for the clinical interpretation of variants vary from disease to disease, in actual practice, experts in each disease area discuss and decide the criteria for clinical significance based on literature information and actual cases in each hospital or institution. For example, in 2015, the American College of Medical Genetics and Genomics (ACMG) and the Association for Molecular Pathology (AMP) published guidelines for the clinical assessment of variants (*1*).

However, the functions of many detected variants from clinical sequence data remain to be elucidated. One of the reasons that variants cannot be interpreted is a lack of functional investigation, since it is unfeasible to validate the pathogenicity of such an enormous number of variants experimentally. As a result, most variants are assigned as being variants of "uncertain significance" (VUS), which means there was no sufficient evidence to judge a relationship between the variant and disease. This is a crucial bottleneck promoting genomic medicine. Therefore, computational methods for predicting the functionality of variants are widely used (*2*), and ACMG-AMP guidelines recommend computational prediction as supporting information for variant interpretation.

A myriad of computational methods have been developed to predict deleterious non-synonymous single nucleotide variants (nsSNVs) (*3*). However, existing methods only utilize features derived from an individual variant or individual gene within the variant. In a physiological setting, genes and their products are associated with each other, and the associations have a strong influence on biological processes and disease development (*4*). These associations can be represented as biological molecular networks, which are composed of diverse biochemical and functional relationships between genes and gene products (*5*), in form of metabolic pathways, signaling pathways, and protein–protein interaction networks. In many cases, human disease will not be caused by individual genomic alterations but by the interaction of individual mutations on multiple other pathways and relationships. It is known that utilizing molecular networks enables the identification of novel genes and pathways associated with a particular disease phenotype (*6*). Nevertheless, there are no computational methods for clinical interpretation of variants using knowledge of biological molecular networks.

Overview of network-based prediction

To address these challenges, in this study, we developed a pathogenicity prediction model with a graph neural network (PathoGN). While existing methods predict the function of variants by machine learning using sequence and structural information of genes and variants, PathoGN achieves highly accurate prediction of variant function (pathogenicity) by learning the structure of a large biological molecular network as background knowledge. As shown in Fig. 1A, PathoGN consists of two major parts: the part that builds a large biomolecular network model by

adding variant data to biological pathways and the part that learns the biomolecular network by deep learning and predicts pathogenicity of the variant.

Table 1 shows the statistics of molecules and variants (nodes) and their connections (edges) that make up the biomolecular network (graph) used in this study. First, we constructed biomolecular networks consisting of proteins and ligands, which include protein–protein interactions, signaling pathways, and metabolic pathways from the Reactome database (*7*), a representative database of biological networks. Then, using the databases with information on effects and pathogenicity of variants, we integrated the relationships (edges) between genes and variants into the Reactome-derived network by considering the same genes and proteins to be common nodes. In addition, the pathogenicity scores predicted by the conventional method were added as nodes of the biomolecular network, resulting in the construction of a large network composed of conventional score-variant-gene-protein-ligand (Fig. 1B).

Next, we used a graph convolutional neural network (GCN), a type of deep learning, to learn the biomolecular network as a knowledge graph and predict the pathogenicity of variants. GCN effectively learns the network structure by using the same concept as convolutional neural network (CNN), which has recently led to a breakthrough in image classification problems (*8–10*). While CNN features a "convolution" process that multiplies and adds weights to the surrounding pixels of a pixel to all the pixels of an input image, GCN can learn the network structure by convoluting the relationships between the nodes of the input network along the network structure. Our approach employs a node-focused prediction (*11*), which predicts a label corresponding to each node. Specifically, we constructed a knowledge graph in which the pathogenicity information of variants registered in the ClinVar database and VariBench datasets is attached as labels to the large molecular network constructed above. By training this graph with GCN, we can predict the pathogenicity or non-pathogenicity of variants with unknown functions, including VUS.

Performance evaluations using a benchmark set

To investigate the effectiveness of utilizing molecular networks and associate information between variants in predicting pathogenicity, we compared the prediction performance of PathoGN and existing popular methods using a benchmark dataset.

Most of the existing methods utilize probabilistic models and machine learning algorithms with features obtained from nucleotide or amino acid sequence conservation and biochemical properties of amino acids that impact protein structure. We compared included MutationTaster2 (*12*), MutationAssessor(*13*), PolyPhen2(*14*), Sorting Intolerant From Tolerant (SIFT)(*15*), likelihood ratio test (LRT)(*16*), and Functional Analysis Through Hidden Markov Models (FATHMM)(*17*), which are commonly used in annotations with clinical interpretation.

Additionally, ensemble methods combine the prediction results from multiple individual methods, improving prediction performance(*18–22*). Thus, we also compared the performance of random forest (RF) and support vector machines (SVM), which were used for ensemble learning with the same predicted scores. RF-based ensemble methods, such as REVEL(*23*) and ClinPred(*24*), generally show excellent performance.

Five benchmark sets by VariBench(*25*) were used for comparison. In the evaluating pathogenicity prediction methods, there was a problem with mixed training and test sets, where

the variant information used to train each method was also included in the evaluation data (called type 1 circularity (*25*)). VariBench provides datasets after removing duplication between datasets in each tool to solve this problem.

Fig. 2A and B show the area under the receiver operating characteristic curve (ROC AUC) and the accuracy for the six existing methods, including Mutation Taster, Mutation Assessor, PolyPhen2, SIFT, LRT, and FATHMM, and other popular algorithms, SVM and RF. For all datasets, the AUC and accuracy of PathoGN significantly outperformed all of the other methods. Notably, PathoGN also showed high accuracy in the SwissVar dataset (Fig. 2A), which contains many unpredictable variants. The difficulty of predicting the SwissVar dataset is thought to be caused by the low ratio of deleterious variants to neutral variants and the mixture of both interpretations in the same gene. PathoGN demonstrated robust accuracies even for unbalanced sets. (See Table S1 and S2 for the details accuracies and Fig. S1 for the ROC curves of all datasets)

In comparison with existing methods and typical algorithms, RF showed the second-best accuracy after PathoGN. Therefore, we examined the difference in prediction results between RF and PathoGN, focusing on genes. The average accuracies for all genes were calculated based on the prediction results for variants of each gene. In Fig. 2C, the comparison between the accuracies of PathoGN and those of RF in the SwissVar dataset is shown. Genes located in the upper left corner in Fig. 2C, which are more accurately predicted by PathoGN, have a higher pathogenic rate of 0 or 1, indicating that they have more similar annotations (benign/pathogenic labels) within the gene. (The comparison between PathoGN and RF for other benchmark datasets are shown in Fig. S2). Many genes were predicted with almost the same accuracies in both PathoGN and RF, but some genes were confirmed to be more accurately predicted by PathoGN (Fig 2E and F). These genes tend to have more than one variant in the same gene and have similar annotations (pathogenic/benign labels) (See also Fig. S3). Since our method enables us to use associated information between variants using the knowledge graph, these variants could be accurately predicted.

Application to ClinVar data

To confirm the applicability of PathoGN using clinically relevant data, we evaluated the performance of PathoGN in predicting the pathogenicity of VUS with the ClinVar dataset. ClinVar is a database of genetic mutation information and its association with diseases and is widely used for interpretation in clinical practice worldwide(*26*). ClinVar updates the clinical significance of each variant when new evidence for interpretation is obtained, providing a dataset that has improved over time. Here, we performed pathogenicity prediction using past datasets of ClinVar and assessed whether these predictions were consistent with the latest annotations.

In the 2020 ClinVar dataset, 908 variants were annotated as pathogenic or benign, but were annotated as belonging to other categories (such as likely pathogenic, likely benign) in the 2019 ClinVar dataset (Fig. 3A). We refer to these variants as “conflict variants.” We performed predictions against the conflict variants to assess how well PathoGN correctly predicts the 2020 labels. PathoGN was trained using 14817 missense variants (pathogenic: 10654, benign: 4163) in the 2019 ClinVar. Then, we used this model to make predictions for conflict variants and verified the model’s accuracy by using the labels of the 2020 dataset as the correct answer set. As a result, PathoGN successfully predicted clinical significance in the 2020 ClinVar dataset

with high predictive performance of accuracy 0.8711 and AUC 0.9402. PathoGN also showed excellent performance with the other past versions of ClinVar (2017, 2018) (Table 2).

In addition, we investigated the significance category of variants for the prediction results in the 2019 ClinVar dataset (Fig. 3B and C). In all clinical significance, PathoGN shows high predictive accuracy. No correlation between label bias or the number of variants within the same gene and accuracy was observed (Fig. S4A and B). On the other hand, the accuracy tended to be lower in categories with fewer variants, such as "Uncertain" and "NonACMG" than in other categories. PathoGN also enabled us to make predictions for variants with “Not provided”. This type of variants tends to comprise those with multiple reported same significance within the same gene (Fig. S4C). This indicates that PathoGN has ability to utilizing neighboring information from the molecular network, such as association information between variants to make these predictions.

Predictions of VUS in the current ClinVar dataset

Next, we looked to identify the clinical significance of variants for which there is currently no known disease association. To do this, we predicted the pathogenicity for all variants that were not annotated as being either pathogenic or benign in the 2020 ClinVar dataset. The model was trained using the labeled data (pathogenic=10,877, benign=7504) and then used to make predictions for the 12,520 unlabeled variants. The details of the variants and prediction results are shown in Fig. 4A. The distribution of predicted pathogenicity probabilities for each significance category is shown in Fig. 4B. The predicted probability of being benign is calculated as one minus the pathogenicity probability. Variants annotated as likely benign in ClinVar tend to have a higher probability of being benign, whereas variants annotated as likely pathogenic tend to have a higher probability of pathogenicity. As a result, 2505 and 1489 variants with uncertain significance were predicted as pathogenic and benign, respectively. The cutoff value for the label was determined by Youden index, which is the most popular approach for determination of cut-point(*27*), and 0.5695 was used. Among them, 1615 and 612 mutations were extracted as the top 10% predicted probability of pathogenic and benign, respectively. Details of these variants and the predicted pathogenicity scores are available in GitHub and MGeND.

Validation of prediction results in the 2020 ClinVar dataset and ACMG classification

In order to evaluate the prediction results for a clinical assessment of VUS (the unlabeled variants), we performed a validation by biocurators assuming a clinical interpretation by an expert panel. Of the predicted unlabeled variants, eight variants for which case information was available from the National Center for Cancer in Japan (NCC) were reviewed for classification according to ACMG guidelines. The clinical genetics professionals in NCC reviewed these variants in conjunction with the available clinical information. The reviewed variants and classification results are described in Fig. 4C.

The ACMG guidelines evaluates variants using multiple criteria and integrates each assessment to determine the variant’s final pathogenicity. The variant’s pathogenicity is represented by five categories: Pathogenic, Likely pathogenic, Likely benign, Benign, and Uncertain significance. These guidelines provide a more objective and accurate assessment.

Variants predicted as “Benign”

Three of the reviewed eight variants were predicted to be "Benign" by PathoGN. Firstly, BRCA c.5660C>T (p.T1887M) is assigned as BP4 because many prediction tools predicted it to be benign/tolerated/neutral. A BP5 was also assigned to this variant because BRCA1 p.E1214X, a pathogenic variant, was concurrently detected in the case. Therefore, although reported in ClinVar with VUS annotations, based on ACMG guidelines, this variant would be determined as Likely benign.

Next, MSH2 c.2533A>G (p.K845E) is annotated as VUS for BS1 and PP3 criteria based on ACMG guidelines. Here, PP3 is predicted to be Pathogenic by multiple algorithms, and BS1 indicates that the allele frequency is higher than expected for the disease. The variant frequency in Human Gene Mutation Database (HGMD) is 4/2,416 (=0.001656) and in ToMMo is 23/9,546 (=0.0025). ToMMo is an integrative Japanese Genome Variation Database of the Japanese population, and the frequencies are obtained from 4.7k Japanese individuals[23]. HGMD manually collects an curates all published germline variants related to human inherited disease (*28*). These frequency values in the general Japanese population present a strong evidence, while the in silico prediction is a supporting evidence in the ACMG guidelines. Taken together, this variant may be considered to be a "benign-like VUS". While many in silico prediction tools assign this variant to be Damaging/Pathogenic, PathoGN predicts "Benign."

The frequency of third variant APC c.3821C>T (p.T1274M) is 3/2,420 (=0.00124) in HGMD and 21/9,546 (=0.0022) in ToMMo, BS1 criteria was also assigned to this variant. Based on the publication search, it was reported that the pathogenic variant had been found in other gene involved in the molecular mechanism of the disease in an case with the same variant (*29*). From this evidence, BP5 was assigned to this variant. Thus, this variant is annotated as "likely benign" based on BS1 and BP5 criteria. These reviews suggest that variants predicted as Benign by PathoGN may be variants that are possible to judge as Benign if more detailed information, such as functional analysis in the population, is obtained in the future.

Variants predicted as "Pathogenic"

Five variants predicted to be "Pathogenic" were also reviewed for ACMG classification. For TP53 c.566 C>T (p.A189V), BS1 and PP3 were assigned, and the classification by the clinical geneticists at NCC is VUS based on the ACMG guidelines. However, the case with this variant is clinically suspected as Li-Fraumeni syndrome. For instance, the mother with breast cancer also had this variant. However, no other segregation data is available, and the biocurators could not assign other ACMG criteria to update the VUS status. The two variants MLH1 c.1153C>T (p.R385C) and MET c.1681C>T (p.P561S) are considered to be VUS with the possibility of benign due to the high allele frequency in the Japanese population. The remaining two variants, CFTR c.4277C>T (p.S1426F) and TSHR c.733G>A (p.G245S), were also reexamined and determined to be VUS, respectively.

Discussion

Our novel pathogenicity prediction approach utilizing molecular networks outperformed existing methods, and the evaluation results indicated the usefulness of molecular network information as a knowledge graph for predicting variant pathogenicity. The predictions made by PathoGN also utilize variant information on the same gene. Thus, variants on genes that have hotspots can be predicted with higher accuracy. The validation using the ClinVar dataset succeeded in accurately predicting the latest clinical significance of variants whose significance was not decided in the

past. The clinical decision of the variant's pathogenicity is comprehensive, considering any information such as the family phenotype information. Therefore, it is not easy to validate the prediction results experimentally. However, PathoGN shows the possibility of making correct predictions for variants whose clinical significance is not clear at present and is expected to accelerate the re-evaluation of VUS.

Moreover, we confirmed the ACMG annotation for the variants that were predicted to be pathogenic or benign. For most of the variants, the guideline-based evidence could not be assigned and clarify their pathogenicity. The current ACMG classification is only designed for high penetrant variants. On the other hand, recent large-scale germline analyses have identified clinically significant variants in the populations that do not exhibit the classic hereditary tumor phenotype. These analyses revealed the need to discuss moderate risk variants, modifier genes, and oligogenic diseases (*30*, *31*). However, the clinical utility for these variants is still undetermined, and are most of them are annotated as VUS.

As a tool to support clinical interpretation, the ability to accurately predict the pathogenicity of high penetrance variants is a prerequisite. On the other hand, the interpretation of VUS in actual genomic medicine requires a more comprehensive judgment based on a thorough understanding of family phenotype and patient information. The data labeled in the current guidelines are not sufficient for this, and a range of additional information needs to be taken into account.

In the future, there is a need to expand the dataset and develop suitable models to achieve this while narrowing down the target diseases. PathoGN is suitable for such an expansion because PathoGN is the knowledge graph-based model. Incorporating detailed information, such as clinical data, into the knowledge graph, PathoGN can be easily expanded to deal with other information. This expansion may help in predictions for variants with unknown functions in the current situation. Additionally, at the moment, PathoGN is only applicable to non-synonymous single nucleotide variants (nsSNV). By improving the input network and knowledge graph, we intend to expand our model to predict insertion/deletion variants, variants in splicing sites, and variants outside of exonic regions.

The era in which individual genomes can be analyzed as a part of the standard health care services will soon be realized, and a vast amount of variants will be identified. We anticipate that the use of molecular networks and knowledge graphs will facilitate these interpretations.

References and Notes

1. S. Richards, N. Aziz, S. Bale, D. Bick, S. Das, J. Gastier-Foster, W. W. Grody, M. Hegde, E. Lyon, E. Spector, K. Voelkerding, H. L. Rehm, Standards and guidelines for the interpretation of sequence variants: A joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.* **17**, 405–424 (2015).

2. R. Karchin, R. Nussinov, Genome Landscapes of Disease: Strategies to Predict the Phenotypic Consequences of Human Germline and Somatic Variation. *PLoS Comput. Biol.* **12**, 2–4 (2016).

3. H. Tang, P. D. Thomas, Tools for Predicting the Functional Impact of Nonsynonymous Genetic Variation. *Genetics*. **203**, 635–647 (2016).

4. T. Hao, Q. Wang, L. Zhao, D. Wu, E. Wang, J. Sun, Analyzing of Molecular Networks for Human Diseases and Drug Discovery. *Curr. Top. Med. Chem.* **18**, 1007–1014 (2018).

5. J. K. Huang, D. E. Carlin, M. K. Yu, W. Zhang, J. F. Kreisberg, P. Tamayo, T. Ideker, Systematic Evaluation of Molecular Networks for Discovery of Disease Genes. *Cell Systems*. **6**, 484-495.e5 (2018).

6. H. Carter, M. Hofree, T. Ideker, Genotype to phenotype via network analysis. *Current Opinion in Genetics and Development* (2013), , doi:10.1016/j.gde.2013.10.003.

7. A. Fabregat, S. Jupe, L. Matthews, K. Sidiropoulos, M. Gillespie, P. Garapati, R. Haw, B. Jassal, F. Korninger, B. May, M. Milacic, C. D. Roca, K. Rothfels, C. Sevilla, V. Shamovsky, S. Shorser, T. Varusai, G. Viteri, J. Weiser, G. Wu, L. Stein, H. Hermjakob, P.

D'Eustachio, The Reactome Pathway Knowledgebase. *Nucleic Acids Res.* **46**, D649–D655 (2017).

8. A. Krizhevsky, I. Sutskever, G. E. Hinton, ImageNet classification with deep convolutional neural networks. *Commun. ACM*. **60**, 84–90 (2017).

9. C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2015), pp. 1–9.

10. Q. V. Le, in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing* (2013), pp. 8595–8598.

11. Z. Zhang, P. Cui, W. Zhu, Deep Learning on Graphs: A Survey. *IEEE Trans. Knowl. Data Eng.*, 1–1 (2020).

12. J. M. Schwarz, C. Rödelsperger, M. Schuelke, D. Seelow, MutationTaster evaluates disease-causing potential of sequence alterations. *Nat. Methods*. **7**, 575–576 (2010).

13. B. Reva, Y. Antipin, C. Sander, Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res.* **39**, e118 (2011).

14. I. A. Adzhubei, S. Schmidt, L. Peshkin, V. E. Ramensky, A. Gerasimova, P. Bork, A. S. Kondrashov, S. R. Sunyaev, A method and server for predicting damaging missense mutations. *Nat. Methods*. **7**, 248–249 (2010).

15. P. Kumar, S. Henikoff, P. C. Ng, Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.* **4**, 1073–1082 (2009).

16. S. Chun, J. C. Fay, Identification of deleterious mutations within three human genomes. *Genome Res.* **19**, 1553–1561 (2009).

17. H. A. Shihab, J. Gough, D. N. Cooper, P. D. Stenson, G. L. A. Barker, K. J. Edwards, I. N. M. Day, T. R. Gaunt, Predicting the Functional, Molecular, and Phenotypic Consequences of Amino Acid Substitutions using Hidden Markov Models. *Hum. Mutat.* **34**, 57–65 (2013).

18. A. González-Pérez, N. López-Bigas, Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. *Am. J. Hum. Genet.* (2011), doi:10.1016/j.ajhg.2011.03.004.

19. D. K. Crockett, P. G. Ridge, A. R. Wilson, E. Lyon, M. S. Williams, S. P. Narus, J. C. Facelli, J. A. Mitchell, Consensus: a framework for evaluation of uncertain gene variants in laboratory test reporting. *Genome Med.* **4**, 48–48 (2012).

20. M. C. Lopes, C. Joyce, G. R. S. Ritchie, S. L. John, F. Cunningham, J. Asimit, E. Zeggini, A Combined Functional Annotation Score for Non-Synonymous Variants. *Hum. Hered.* **73**, 47–51 (2012).

21. M.-X. Li, H.-S. Gui, J. S. H. Kwan, S.-Y. Bao, P. C. Sham, A comprehensive framework for prioritizing variants in exome sequencing studies of Mendelian diseases. *Nucleic Acids Res.* **40**, e53–e53 (2012).

22. J. Bendl, J. Stourac, O. Salanda, A. Pavelka, E. D. Wieben, J. Zendulka, J. Brezovsky, J. Damborsky, PredictSNP: Robust and Accurate Consensus Classifier for Prediction of Disease-Related Mutations. *PLoS Comput. Biol.* **10**, e1003440–e1003440 (2014).

23. N. M. Ioannidis, J. H. Rothstein, V. Pejaver, S. Middha, S. K. McDonnell, S. Baheti, A. Musolf, Q. Li, E. Holzinger, D. Karyadi, L. A. Cannon-Albright, C. C. Teerlink, J. L. Stanford, W. B. Isaacs, J. Xu, K. A. Cooney, E. M. Lange, J. Schleutker, J. D. Carpten, I. J. Powell, O. Cussenot, G. Cancel-Tassin, G. G. Giles, R. J. MacInnis, C. Maier, C.-L. Hsieh, F. Wiklund, W. J. Catalona, W. D. Foulkes, D. Mandal, R. A. Eeles, Z. Kote-Jarai, C. D. Bustamante, D. J. Schaid, T. Hastie, E. A. Ostrander, J. E. Bailey-Wilson, P. Radivojac, S. N. Thibodeau, A. S. Whittemore, W. Sieh, REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. *Am. J. Hum. Genet.* **99**, 877–885 (2016).

24. N. Alirezaie, K. D. Kernohan, T. Hartley, J. Majewski, T. D. Hocking, ClinPred: Prediction Tool to Identify Disease-Relevant Nonsynonymous Single-Nucleotide Variants. *Am. J. Hum. Genet.* **103**, 474–483 (2018).

25. D. G. Grimm, C. A. Azencott, F. Aicheler, U. Gieraths, D. G. Macarthur, K. E. Samocha, D. N. Cooper, P. D. Stenson, M. J. Daly, J. W. Smoller, L. E. Duncan, K. M. Borgwardt, The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity. *Hum. Mutat.* **36**, 513–523 (2015).

26. M. J. Landrum, J. M. Lee, M. Benson, G. Brown, C. Chao, S. Chitipiralla, B. Gu, J. Hart, D. Hoffman, J. Hoover, W. Jang, K. Katz, M. Ovetsky, G. Riley, A. Sethi, R. Tully, R. Villamarin-Salomon, W. Rubinstein, D. R. Maglott, ClinVar: Public archive of interpretations of clinically relevant variants. *Nucleic Acids Res.* **44**, D862–D868 (2016).

27. R. Fluss, D. Faraggi, B. Reiser, Estimation of the Youden Index and its associated cutoff point. *Biom. J.* **47**, 458–472 (2005).

28. P. D. Stenson, E. V. Ball, M. Mort, A. D. Phillips, J. A. Shiel, N. S. T. Thomas, S. Abeysinghe, M. Krawczak, D. N. Cooper, Human Gene Mutation Database (HGMD): 2003 update. *Hum. Mutat.* **21**, 577–581 (2003).

29. B. Kim, D. Won, M. Jang, H. Kim, J. R. Choi, T. I. Kim, S.-T. Lee, Next-generation sequencing with comprehensive bioinformatics analysis facilitates somatic mosaic APC

gene mutation detection in patients with familial adenomatous polyposis. *BMC Med. Genomics*. **12**, 103 (2019).

30. A. J. Marian, Clinical Interpretation and Management of Genetic Variants. *JACC Basic Transl Sci*. **5**, 1029–1042 (2020).

31. J. Taeubner, D. Wieczorek, L. Yasin, T. Brozou, A. Borkhardt, M. Kuhlen, Penetrance and Expressivity in Inherited Cancer Predisposing Syndromes. *Trends Cancer Res.* **4**, 718–728 (2018).

Acknowledgments:

Funding: This work was supported by the Japan Agency for Medical Research and Development (AMED) under grant number JP18kk0205013h0003. This work was supported by MEXT as "Program for Promoting Researches on the Supercomputer Fugaku" (Application of Molecular Dynamics Simulation to Precision Medicine Using Big Data Integration System for Drug Discovery).

Author contributions:

Conceptualization: YO, MK, RK

Methodology: RK, MK, AT, YO

Investigation: RK, MK, AT, MN, TY, MH, NT

Visualization: MK, YT

Writing – original draft: MK, YO, AT, RK

Writing – review & editing: All

Competing interests: The authors have no competing interests.

Data and materials availability:

All data are available in the main text or the supplementary materials. The knowledge graph used for PathoGN was constructed from public databases; Reactome, VariBench, ClinVar. The open-source Python code to run the demo with PathoGN and the pathogenicity scores for ClinVar 2020 dataset predicted by PathoGN are available on GitHub (https://github.com/clinfo/PathoGN.git). The pathogenicity scores for ClinVar 2020 dataset are also available in MGeND (https://mgend.med.kyoto-u.ac.jp/).

Supplementary Materials

Materials and Methods

Figs. S1 to S4

Tables S1 to S6

References (*1–12*)

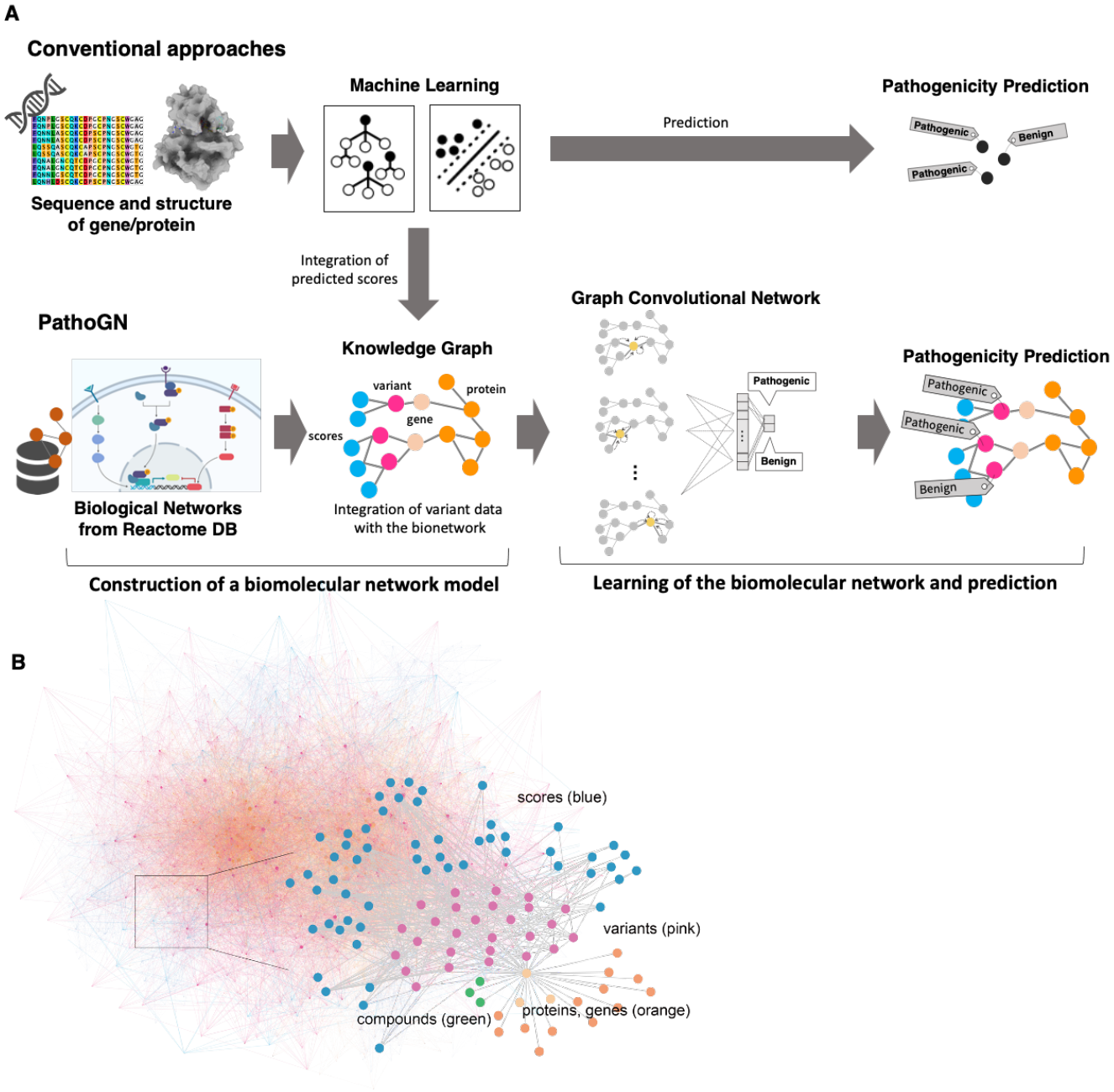


Fig. 1. Schematic graph-based prediction method of PathoGN (**A**) Overview of graph-based prediction. Conventional approaches use the information for individual variants such as that obtained from genome sequences. By contrast, PathoGN utilizes biological networks represented in a graph structure. Combining biological network information, disease-associated variants information, and prediction scores by conventional tools, a knowledge graph is constructed and used as an input for graph convolutional networks (GCN). Proteins, genes, variants, compounds, and prediction scores are represented by a node, and relationships between them are denoted as edges on a graph. (**B**) Knowledge graph image as input of PathoGN. The statistical detail of graph is shown in Table 1.

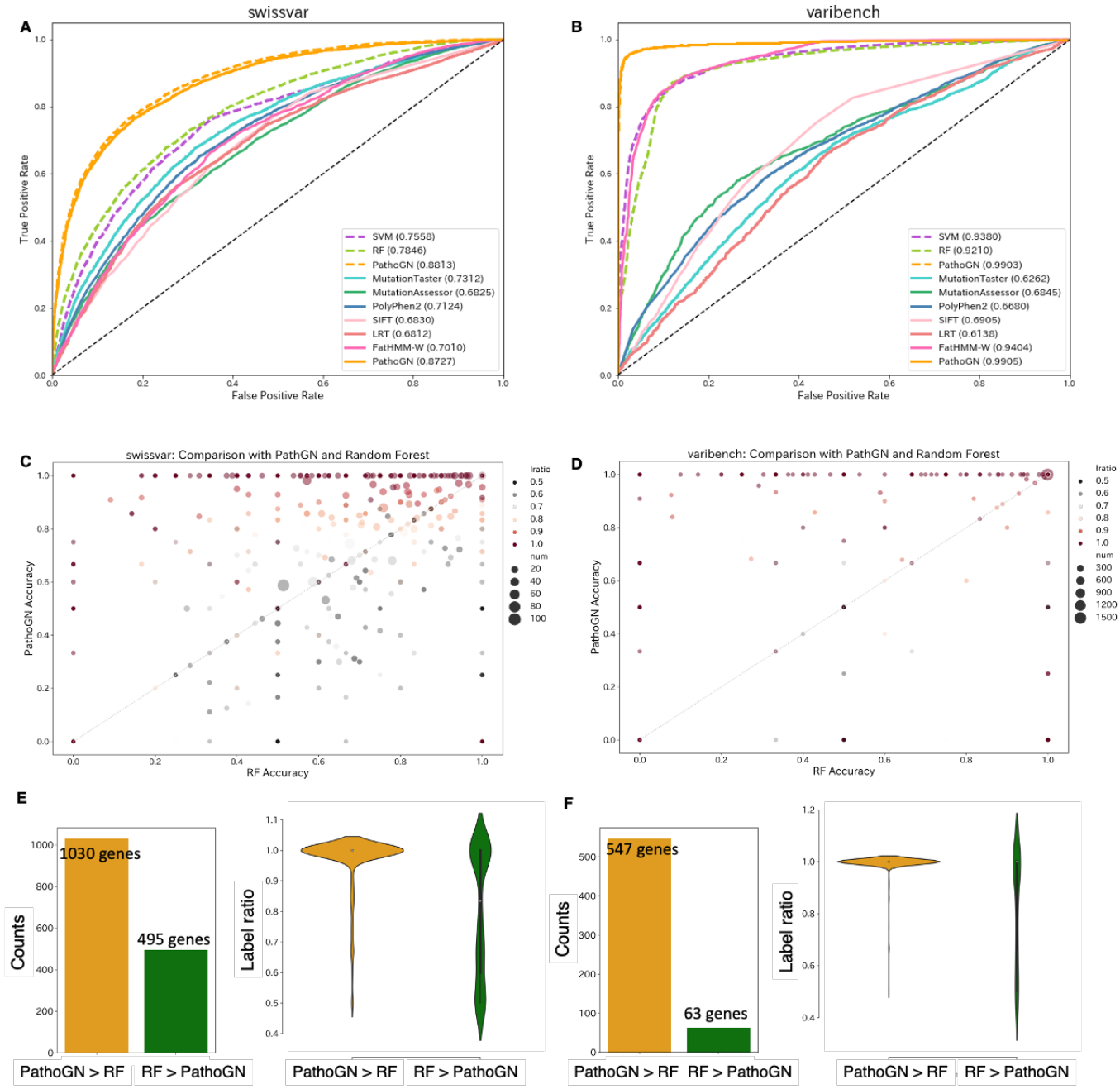


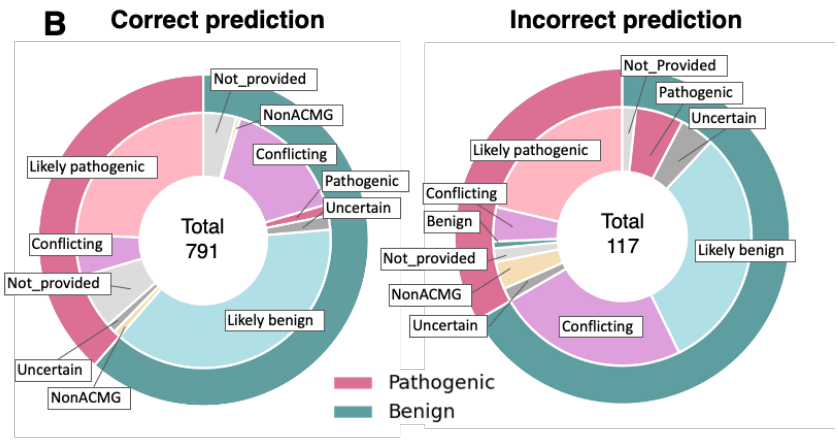
Fig. 2. Performance evaluation. (**A, B**) ROC curves for our method (PathoGN), other non-ensemble methods, and two popular algorithms (SVM and RF) on SwissVar dataset and Varibench dataset. Solid line indicates ROC curves for PathoGN and other non-ensemble methods. For this plot, only variants for which predicted scores by all comparator tools could be obtained were used in the evaluation. Dashed line indicates ROC curve for PathoGN and other algorithms using all variants in the dataset. (**C**, **D**) Accuracy comparison with our approach (PathoGN) and Random Forest (RF) on SwissVar dataset and Varibench dataset. Each point represents an aggregated accuracy (mean accuracy) for each gene. The dot size represents the number of variants present in each gene in the dataset; the larger dot, the more variants on the

gene are included in the dataset. The color indicates the bias of labels (label ratio) in a gene. For example, the red color indicates that all variants present in the gene have the same labels. (**E, F**) Gene counts and the label bias for PathoGN and RF. Gene were classified according to which method was more accurate (high accuracy value), and the number of genes and the bias of the labels were compared. Bar charts show the number of genes, violinplots indicate the label ratio for each gene. Orange is for genes for which prediction by PathoGN was more accurate, and green is for genes for which prediction by RF was more accurate.

A

2019 ClinVar Label	2020 ClinVar Label: Pathogenic	2020 ClinVar Label: Benign
Pathogenic	0	18
Likely_pathogenic	217	0
Benign	1	0
Likely_benign	0	335
Conflicting	47	153
Uncertain	11	19
NonACMG	11	4
Not_provided	57	35

B



C

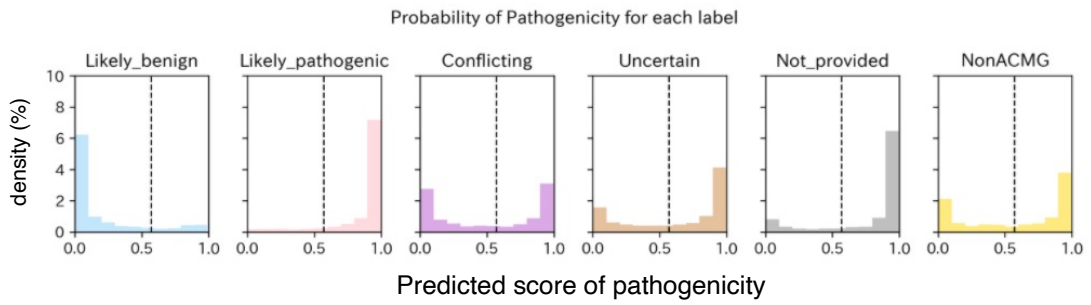
	PathoGN prediction: Pathogenic Acc	PathoGN prediction: Benign Acc	PathoGN prediction: All Acc
Pathogenic	--	0.6111	0.6111
Likely_pathogenic	0.8848	--	0.8848
Benign	0.0000	--	0.0000
Likely_benign	--	0.8925	0.8925
Conflicting	0.8936	0.8170	0.8350
Uncertain	0.8182	0.7368	0.7667
NonACMG	0.6364	1.0000	0.7333
Not_provided	0.9649	0.9429	0.9565

Figure 3. Details of predictions made with the ClinVar dataset. (**A**) Details of labels in the ClinVar dataset. Each row indicates variant's labels in the 2019 ClinVar dataset, and the columns indicate labels assigned to the variants in the 2020 ClinVar dataset. (**B**) Each pie chart shows the category of variants. The left pie chart shows distributions of variants predicted correctly, the right pie chart shows those variants that were predicted incorrectly. The outer pie indicates the labels in the 2020 ClinVar dataset, and the inner pie shows the labels in the 2019 ClinVar dataset. (**C**) Accuracy for each significance category of variants. Pathogenic/Benign Acc column indicates prediction accuracy for the variants annotated with pathogenic/benign in 2020 ClinVar, and All Acc columns total prediction accuracy for the variants in each category.

A

Clinical label in 2020 ClinVar		Predicted label	
Type	total	Pathogenic	Benign
Likely benign	1027	142	885
Likely pathogenic	2238	1995	243
Conflicting	2465	1210	1255
Uncertain	3994	2505	1489
Not provided	2077	1677	400
NonACMG	719	419	300

B



C

Variants				ACMG Evidence satisfied																												variant class	Clinical comments	GCN Prediction
Gene	Reference	cDNA	protein	PVS1	PS1	PS2	PS3	PS4	PM1	PM2	PM3	PM4	PM5	PM6	PP1	PP2	PP3	PP4	PP5	BA1	BS1	BS2	BS3	BS4	BP1	BP2	BP3	BP4	BP5	BP6	BP7			
BRCA2	NM_000059.3	c.5660C>T	p.T1887M																										■	■			Likely benign	Likely benign	B
MHS2	NM_000251.2	c.2533A>G	p.K845E														■				■												VUS	benign-like VUS	B
APC	NM_001127511.2	c.3821C>T	p.T1274M																		■								■				Likely benign	Likely benign	B
TSHR	NM_000369.2	c.733G>A	p.G245S														■				■												VUS	VUS	P
TP53	NM_000546.5	c.566C>T	p.A189V														■				■												VUS	VUS	P
MLH1	NM_000249	c.1153C>T	p.R385C														■				■								■				VUS	benign-like VUS	P
CFTR	NM_000492.3	c.4277C>T	p.S1426F														■		■														VUS	VUS	P
MET	NM_001324402.1	c.1681C>T	p.P561S																		■												VUS	benign-like VUS	P

Fig. 4. Validating predictions using ClinVar dataset. (**A**) Details of the 2020 ClinVar dataset. The "Total" column represents the number of variants included in the 2020 ClinVar dataset and used in the prediction. Variants for which a score was available from Varibench were used in the prediction. The "Predicted label" columns show the number of labels predicted by PathoGN. (**B**) For each significance label, the distribution of the predicted pathogenicity probability is shown. The vertical axis shows the frequency percentage normalized by the number of variants with each label, and the horizontal axis shows the predicted pathogenicity probability. Here, the predicted probability of Benign is equal to the pathogenicity probability minus one. Dash line indicates cut-off value for the label (Youden index = 0.5695) (**C**) Details of the variants reviewed with ACMG guideline. All variants are annotated as “Uncertain significance” in 2020 ClinVar dataset.

Table 1. The statistics of knowledge graph. (**A**) The number of nodes and edges in the knowledge graph constructed from the molecular network (Reactome). Reactome instance is a Reactome record, which indicates a reaction by any of the components in the Reactome database. (**B**) The number of nodes and edges in the knowledge graph constructed each benchmark dataset and molecular network graph. The "total" edges include self-edges and bidirectional edges, and the "unique" columns is calculated by (total edges – the number of nodes)/2.

	nodes			edges		
	Protein	Compound	Reactome instance	Protein - Protein	Protein - Compound	Protein-Reactome
Basic biological network	5401	248	4155	7270	1138	8394

	nodes					edges				
dataset	total	variant	gene	proteins	scores	total	unique	score-variant	gene-variant	protein-variant
exovar	24865	8850	7739	3750	123	254067	114601	141600	17700	53100
humvar	66068	40389	11774	9376	126	1052206	493069	646224	80778	242334
predictSNP	30794	16098	9614	552	127	433948	201577	257568	32196	96588
swissvar	30447	12729	8473	4718	124	352745	161149	203664	25458	76374
varibench	27358	10266	8216	4348	125	290544	131593	164256	20532	61596

Table 2. Details of the ClinVar dataset and performance results. Training set includes variants annotated as Pathogenic or Benign in ClinVar. For the "Infer" set, variants have conflicting clinical significance among 2019, 2018, 2017, and 2020. The clinical significance in 2020 was assumed to be the correct label, and the accuracy and AUC were calculated accordingly.

dataset	Training set		Infer set		Accuracy	AUC
	pathogenic	benign	pathogenic	benign		
2019	10654	4163	344	564	0.8623	0.9405
2018	11041	3893	409	727	0.8680	0.9349
2017	10490	3555	555	836	0.8742	0.9465