Computational and Structural Biotechnology Journal 23 (2024) 843–858

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Research article

Large language models assisted multi-effect variants mining on cerebral cavernous malformation familial whole genome sequencing

Yiqi Wang [a,b,d], Jinmei Zuo [e], Chao Duan [a,b], Hao Peng [b,c], Jia Huang [f], Liang Zhao [d,*], Li Zhang [b,c,**], Zhiqiang Dong [a,b,***]

[a] College of Biomedicine and Health, College of Life Science and Technology, Huazhong Agricultural University, No.1, Shizishan Street, Wuhan 430070, Hubei, China
[b] Center for Neurological Disease Research, Taihe Hospital, Hubei University of Medicine, No.32, Renmin South Road, Shiyan 442000, Hubei, China
[c] Department of Neurosurgery, Taihe Hospital, Hubei University of Medicine, No.32, Renmin South Road, Shiyan 442000, Hubei, China
[d] Precision Medicine Research Center, Taihe Hospital, Hubei University of Medicine, No. 32, Renmin South Road, Shiyan 442000, Hubei, China
[e] Physical Examination Center, Taihe Hospital, Hubei University of Medicine, No. 32, Renmin South Road, Shiyan 442000, Hubei, China
[f] The Second Clinical Medical College, Lanzhou University, No. 222, South Tianshui Road, Lanzhou 730030, Gansu, China

ARTICLE INFO

Keywords:
Whole genome sequencing
Cerebral cavernous malformation
Deep learning
Large language model
Natural language processing

ABSTRACT

Cerebral cavernous malformation (CCM) is a polygenic disease with intricate genetic interactions contributing to quantitative pathogenesis across multiple factors. The principal pathogenic genes of CCM, specifically KRIT1, CCM2, and PDCD10, have been reported, accompanied by a growing wealth of genetic data related to mutations. Furthermore, numerous other molecules associated with CCM have been unearthed. However, tackling such massive volumes of unstructured data remains challenging until the advent of advanced large language models. In this study, we developed an automated analytical pipeline specialized in single nucleotide variants (SNVs) related biomedical text analysis called BRLM. To facilitate this, BioBERT was employed to vectorize the rich information of SNVs, while a deep residue network was used to discriminate the classes of the SNVs. BRLM was initially constructed on mutations from 12 different types of TCGA cancers, achieving an accuracy exceeding 99%. It was further examined for CCM mutations in familial sequencing data analysis, highlighting an upstream master regulator gene fibroblast growth factor 1 (FGF1). With multi-omics characterization and validation in biological function, FGF1 demonstrated to play a significant role in the development of CCMs, which proved the effectiveness of our model. The BRLM web server is available at http://1.117.230.196.

1. Introduction

Cerebral cavernous malformation (CCM; OMIM 116860), also known as cerebral cavernous angiomas, can manifest as sporadic or autosomal dominant conditions. These conditions consist of a varied range of relatively prevalent lesions that have important clinical implications [58]. These angiomas may arise sporadically or be inherited, with identified causative genes primarily attributed to KRIT1, CCM2, and PDCD10 [31], which have been linked to the molecular diagnostic criteria of the condition for the last two decades [44]. However, not all sources succeeded in identifying mutations within specified loci of the above three genes. The advent of next-generation sequencing techniques has unveiled a broader spectrum of CCM-associated genes. With regard to transcriptome sequencing, the focus had been on 1325 genes displaying differential expression between CCM endothelial cells (CCMECs) and Human brain microvascular endothelial cells (HBMECs)

Abbreviations: CCM, Cerebral cavernous malformation; WGS, Whole genome sequencing; SNVs, Single nucleotide variants; FGF, Fibroblast growth factor; LLM, Large language model; NLP, Natural language processing; ACMG, The American College of Medical Genetics and Genomics; VQSR, Variant quality score recalibration; PPI, Protein-Protein Interaction; PMAP, Pathway mutations accumulative perturbation; GO, Gene Ontology; KEGG, Kyoto Encyclopedia of Genes and Genomes; TCGA, the Cancer Genome Atlas; UMAP, Uniform Manifold Approximation and Projection.

* Corresponding author.
** Corresponding author at: Center for Neurological Disease Research, Taihe Hospital, Hubei University of Medicine, No.32, Renmin South Road, Shiyan 442000, Hubei, China.
*** Corresponding author at: College of Biomedicine and Health, College of Life Science and Technology, Huazhong Agricultural University, No.1, Shizishan Street, Wuhan 430070, Hubei, China.
E-mail addresses: s080011@e.ntu.edu.sg (L. Zhao), zhanglith@163.com (L. Zhang), dongz@mail.hzau.edu.cn (Z. Dong).

https://doi.org/10.1016/j.csbj.2024.01.014
Received 16 October 2023; Received in revised form 4 January 2024; Accepted 19 January 2024
Available online 1 February 2024

[44], along with 80 enriched pathway terms [31]. Additionally, whole genome sequencing (WGS) had even detected instances of heterozygous inversion without mutations [47]. Presently, the most advanced approach, single cell RNA sequencing (scRNA-seq), has provided a comprehensive gene expression atlas for mouse models of CCM across various cell groups [36].

Despite the collective findings from these aforementioned studies, an all-encompassing pathogenic genetic mutations for CCM remains elusive. Consequently, the substantial body of recent research has shifted its focus from singular genes to polygenic factors associated with diseases. This study aims to explore the genetic regulatory aspects of inherited CCM WGS data, based on SNVs' annotated textual information mining.

Those multi-factor texts were presented in various forms, encompassing functional descriptions [58], regulatory networks [7], scientific literatures [54], and retrievable databases [38], thereby rendering these discoveries as distinct "informational isolated island" [17]. Complex disease traits have been revealed to stem from the interplay of multiple elements. Among the potential contributors to the pathogenesis of complex diseases, SNVs, genes, and gene interactions stand out as prominent factors [50]. In response, researchers have formulated various algorithms aimed at scrutinizing biomarkers. Examples include the penalized logistic regression [8], multi-factor dimension reduction [8], set association [59], Bayesian network and random forest method [62], to name a few. However, these algorithms are limited by the need for explicit feature extractions, which must be engineered artificially.

While genetic information is inherently an aspect of natural language, where large language model (LLM) had been devoted to natural language processing (NLP) in comprehensive feature tracking and summarization. The two primary NLP technologies currently utilized are GPT [55] and BERT [13]. GPT has not received any specific biological training, although the recently developed scGPT [11] had been limited to single cell sequencing analysis. BioBERT [25], on the other hand, specializes in the biomedical field and has been utilized for biomedical natural language encoding. Such language learning model had been applied to molecular interactions mining [48].

By using BioBERT [25] for NLP, unstructured annotation texts were converted into per-SNV vectors. All SNV information was transformed into computationally manageable vectors, which was a great challenge to overcome gradient disappearance and performance degradation in traditional deep neural networks. These issues were effectively addressed since the invention of ResNet [51]. The ResNet50 was borrowed in this study for BioBERT encoded vector classification with intention, as it provides an ideal balance of depth and computational efficiency for a variety of tasks [51]. Thus, classification of input vectors for vast variants was carried out using reconstructed ResNet50 [51], although ResNet was originally employed in image recognition.

Using BioBERT [25] as encorder and ResNet50 [51] as classifier, we name our model as BRLM, short for BioBERT vectorized input for ResNet classification language model, dedicated to pathogenicity classification of SNVs in annotation texts. BRLM's performance was validated on 12 TCGA cancers to ensure its accuracy and robustness. To demonstrate its proficiency in resolving practical problems, we successfully classified SNVs for familial CCM WGS variants, and analyzed mutated genes involved in perturbated pathways.

BRLM was ultimately applied on CCM risk element mining to isolate the top three risk levels SNVs (pathogenic, likely pathogenic and uncertain significance) for further investigation. Three-level SNVs were verified by genetic functional domains and protein-protein interactions (PPIs) to demonstrate its effectiveness regarding pathogenicity. Subsequently, these three-class mutated genes were undergo KEGG pathway enrichment analysis with up- and downstream cumulative effect evaluation. The integrative results outlined an upstream regulator gene FGF1, which provided a clear and concise multi-omics atlas of CCM functional landscape.

2. Results

2.1. Accuracy evaluation of BRLM

The BRLM model was initially trained on 12 TCGA cancers, encompassing 367,224 SNVs from 3104 patients. The datasets sourced from TCGA were diverse, containing mutations from various organs and volumes to achieve optimal parameters for best performance. The dataset comprised a large-scale cases of popular cancers (ACC, BRCA, and GBM), a small-scale cases of rare cancers (CHOL, DLBC, and KICH), as well as a medium-sized cases of common cancers. Detailed information about the 12 TCGA datasets is presented in Table 1, including descriptions, patient numbers, mutation amounts, and links for each dataset. The density distribution of the aforementioned SNVs is depicted in Fig. 1A (left).

Concurrently, the vectors were input into the ResNet50 classifier, whose structure is illustrated in the mid-panel of Fig. 1A along with its classification results in the right panel. Adhering to the American College of Medical Genetics and Genomics (ACMG) [39] criteria for variant

Table 1
TCGA data information.

Abbreviation	Cancer	Patients	SNVs	Link
ACC	Adrenocortical carcinoma	90	20,161	https://portal.gdc.cancer.gov/projects/TCGA-ACC
BLCA	Bladder Urothelial Carcinoma	130	39,309	https://portal.gdc.cancer.gov/projects/TCGA-BLCA
BRCA	Breast invasive carcinoma	982	84,713	https://portal.gdc.cancer.gov/projects/TCGA-BRCA
CESC	Cervical squamous cell carcinoma and endocervical adenocarcinoma	194	35,606	https://portal.gdc.cancer.gov/projects/TCGA-CESC
CHOL	Cholangiocarcinoma	35	3833	https://portal.gdc.cancer.gov/projects/TCGA-CHOL
COAD	Colon adenocarcinoma	154	54,349	https://portal.gdc.cancer.gov/projects/TCGA-COAD
DLBC	Lymphoid Neoplasm Diffuse Large B cell Lymphoma	48	7276	https://portal.gdc.cancer.gov/projects/TCGA-DLBC
ESCA	Esophageal carcinoma	185	36,288	https://portal.gdc.cancer.gov/projects/TCGA-ESCA
GBM	Glioblastoma multiforme	290	22,044	https://portal.gdc.cancer.gov/projects/TCGA-GBM
KICH	Kidney Chromophobe	66	5923	https://portal.gdc.cancer.gov/projects/TCGA-KICH
KIPAN	KICH (Kidney Chromophobe), KIRC (Kidney renal clear cell carcinoma), and KIRP (Kidney renal papillary cell carcinoma)	644	47,875	https://www.linkedomics.org/data_download/TCGA-KIPAN/
LGG	Brain Lower Grade Glioma	286	9847	https://portal.gdc.cancer.gov/projects/TCGA-LGG

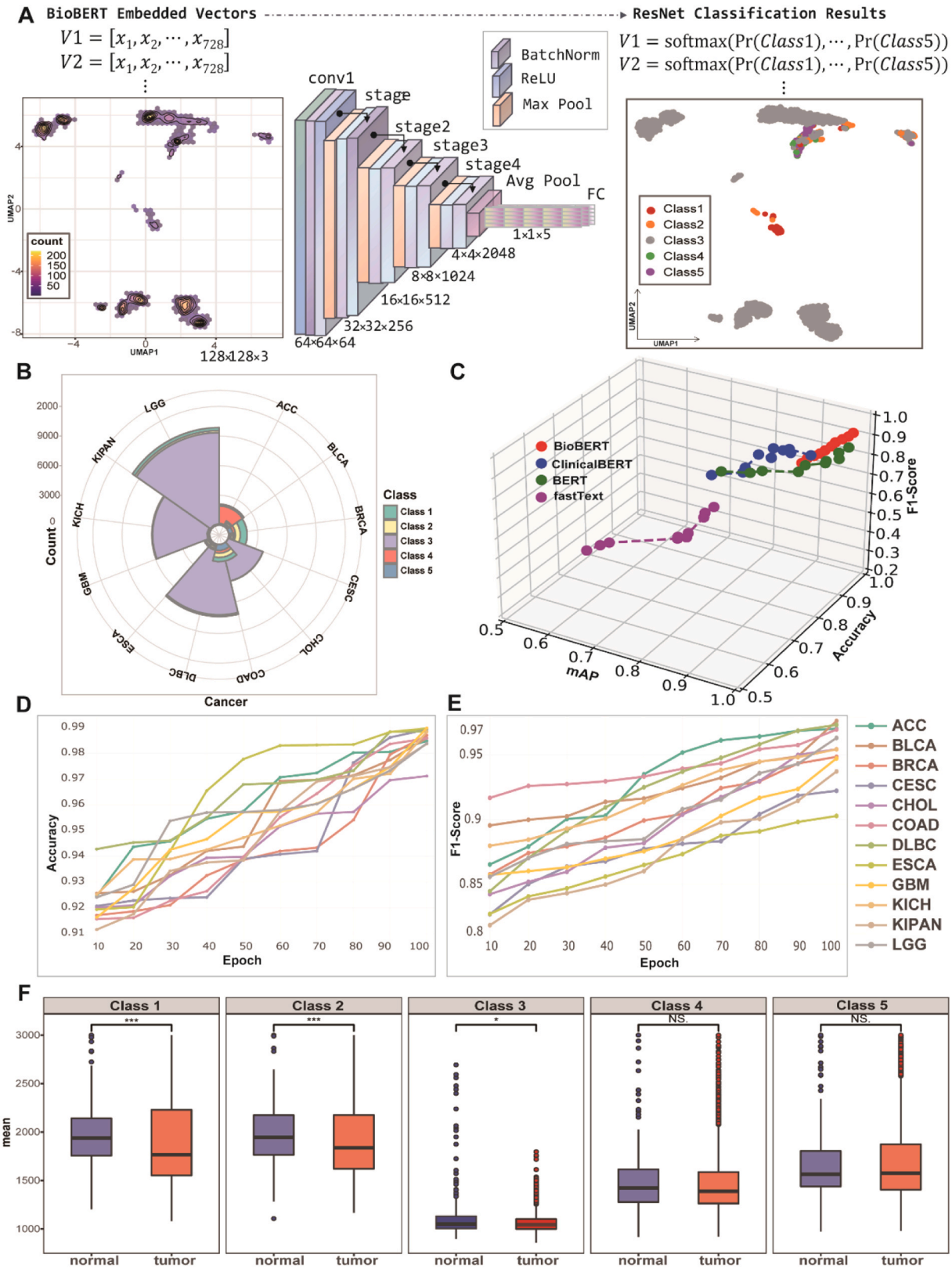

Fig. 1. BRLM model construction and performance evaluation. (A)BRLM structure, including BioBERT encoded annotation of SNVs into 728-dimensional vectors visualized by UMAP (left), ResNet50 model architecture for SNV classification (middle); and classification results presented by UMAP (right). (B) TCGA pan-cancer classified variants distribution in Nightingale rose diagram. (C) Classification performances among four biomedical encoders in 12 TCGA cancers after 100 epochs. (D) Classification accuracy of BRLM per 10 epochs in 12 TCGA cancers. (E) Classification F1-score of BRLM per 10 epochs in 12 TCGA cancers. (F) Expression comparison between tumor and normal tissues as validation of TCGA variants classification.

classification, all TCGA SNVs were categorized into five classes, namely Class 1 (pathogenic), Class 2 (likely pathogenic), Class 3 (uncertain significance), Class 4 (likely benign), and Class 5 (benign). Likewise, the final classification results are highly consistent with the clusters formed by UMAP. It is notable that mutations in Class 3 (uncertain significance) comprised the largest fraction of the sequencing analysis outcomes, and the results of BRLM align closely with this distribution.

In addition to presenting overall statistics, we analyzed the distribution of classes within each tumor type, as depicted in Fig. 1B. Due to the diverse nature of tumor variants, their classification proportions exhibit distinct characteristic. Class 3, representing SNVs of uncertain significance, is prominent across multiple tumors (CESC, CHOL, DLBC, GBM, KICH, KIPAN, and LGG). Conversely, for other tumors such as BRCA, BLCA, COAD, and ESCA, the classes demonstrate relatively equal proportions. Notably, more than 90% of the cases in ACC are constituted by Class 4 (likely benign) variants. These findings are consistent with pre-established pathogenic evidence, that certain types of cancers such as ACC, BRCA, CESC, and COAD may exhibit a focal concentration of pathogenic SNVs, resulting in clear categorized mutations and drug targets. However, most carcinomas exhibit significant heterogeneity and individual variability, making it difficult to classify intricate risk factors.

After completing 100 training epochs across 12 cancers in TCGA, the performance comparison of frequently used encoders in the biomedical field (BERT, ClinicalBERT, and fastText) is presented in Fig. 1C, which is a three-dimensional line graph for the accuracy, mAP (mean average precision), and F1-score. The performance reveals that BioBERT is superior to other encoders, particularly in accuracy and F1-score.

To further evaluate the model's performance, we presented the curve graphs for accuracy and F1-score at 10-epoch intervals for each type of cancer in Fig. 1D and E. The plots demonstrate consistent testing performance throughout each training epoch, with the optimal values of accuracy ranging from 0.91 to 0.99 and F1-score ranging from 0.80 to 0.97. As the F1-score represents the harmonic mean of precision and recall, their line graphs were further included in Supplementary Figure S1, where the optimal precision ranges from 0.821 to 0.988 and the optimal recall ranges from 0.753 to 0.967 across the 12 datasets.

To further validate the classification results, we extracted tumor-related variants and integrated them with TCGA RNAseq dataset expression analysis. Differential expression comparison between tumor and normal conditions for each class is illustrated in Fig. 1F. The classified cancer-related SNVs (Class 1 and 2) exhibit extremely-significant differences, while Class 3 (uncertain disease-associated probability) demonstrates significant difference. In contrast, no significant expression differences are observed between the normal and tumor groups for neutral mutations (class 4 and 5).

2.2. Generalizability assessment of BRLM

Following the optimal model structure determined through multi-cancer verification on the TCGA dataset, BRLM was subsequently employed to analyze a familial WGS data with the aim of identifying variants associated with CCM. The genetic pedigree of the family was composed of four CCM affected individuals and four unaffected individuals marked by asterisks; see Supplementary Figure S2 for details. Based on the WGS analysis with Genome Analysis Toolkit (GATK) [5] workflow for germline short variants, 409,666 germline variants were identified. Several benign SNVs were found in the three well-known CCM-related genes (KRIT1, CCM2 and PDCD10), while a large amount of variants with unknown impacts were also identified. These benign mutations have a high frequency in population and are located within intronic areas, whose detailed information can be found in Supplementary Table S1. We thus developed BRLM to assist the labor-intensive tasks of SNVs interpretation and integration. The entire variants were then annotated and transformed into BioBERT embedded vectors.

Classification results from ResNet50 are visualized in Fig. 2A using the UMAP technique, which underscores the specificity of BRLM in SNV classification through clusters of aggregated distributions. Validations of these results drawn upon both SIFT scores and Clinvar records are demonstrated in Fig. 2B and C. The SIFT scoring mechanism ranges from 0 to 1 (Fig. 2B), with color intensity reflecting the severity of detrimental effects where a score of 0 indicates the most deleterious mutation. On the other hand, Fig. 2C illustrates the Clinvar category mapping. Based on the comparison between the two plots, BRLM shows a good performance in identifying CCM-related variants. Specifically, Class 1 and 2 in Fig. 2A are consistent with risk SNVs as aligned with the yellow clusters in Fig. 2B and the red unlabeled variants in Fig. 2C. Similarly, the classification results for Class 4 and 5 also correlate well with the patterns observed in both graphs.

However, the classification of variants with uncertain significance (Class 3) remains a contentious issue due to discrepancies between the score deficiency highlighted in Fig. 2B and the benign or likely benign categorization in Fig. 2C. To further assess the effectiveness of our classification results, an additional eleven algorithms for variant classification were utilized. The detailed UMAP plots of the outcomes can be found in Supplementary Figure S3. Of the total eleven algorithms, nine could only annotate a limited number of variants albeit with diverse clusters, whereas the remaining two were incapable of categorizing the variants. The majority of classified SNVs demonstrated to be consistent with ours. Additionally, BRLM was able to estimate unpredicted variants by other algorithms, highlighting the comprehensiveness and superiority of our model.

Owing to the doubtable SNVs in Class 3, we investigated their functional regions. In Fig. 2D, statistical data is presented regarding overall functional variants (left) and specific exonic variants (right). Due to the substantial number of functional mutations within Class 3, we further localized their chromosome distribution in Fig. 2E, corresponding to functional and specific exonic SNVs. The overall distribution pattern remains consistent across both sets of variants. The mutation coordinate and distribution is an important part of the input text. However, functional mutations are distributed near exonic mutations, revealing a possible regulatory impact.

To narrow down valid data, we extracted positive sites from SNVs depicted in Fig. 2E. These positive sites were defined by the variant quality score recalibration (VQSR) model in GATK. VQSR learned features by machine-learning algorithms from mutation sites to distinguish positives from negatives. The detailed parameter settings for VQSR are introduced in Section 3.3.

Circos graphs presented in Fig. 2F and G illustrate the positive sites distribution comparison between Class 1, 2 and 3. Fig. 2F portrays greater risk variants within Class 1 and 2, whereas Fig. 2G emphasizes abundant mutual effects within Class 3. Mutated genetic functional regions are colored at the outermost chromosomes, where exonic variants are highlighted in red. Their frequencies are depicted in curves and bars of the two middle rings that are obtained from the 1000 Genomes and the Genome Aggregation Database (gnomAD), and internal lines indicate Protein-Protein Interactions (PPI) among them. Conversely, deleterious interactions among genes in Class 1 and 2 are only centered on three links with six genes highlighted in bold: FGF1, FGF6, ABCB1, ABCG2, IL4R, and CTLA4. Consequently, interactive effects among these genes have higher priorities in our following analysis.

2.3. Enrichment analysis of candidate genes and pathways

To explore the biological functions affected by the categorized CCM-related mutations, an initial KEGG pathway enrichment analysis was performed for the mutated genes within Class 1, 2 and 3, yielding 661 enriched pathways. The simplifyEnrichment package was then invoked to cluster the similarity matrix of enriched terms into groups using "binary cut" [16], and the results are illustrated by the heatmap in Fig. 3A. Based on the eight clusters with descriptive vocabulary frequencies indicated by font size, the mutated genes play significant roles in the developmental processes and signaling activities on cell adhesion

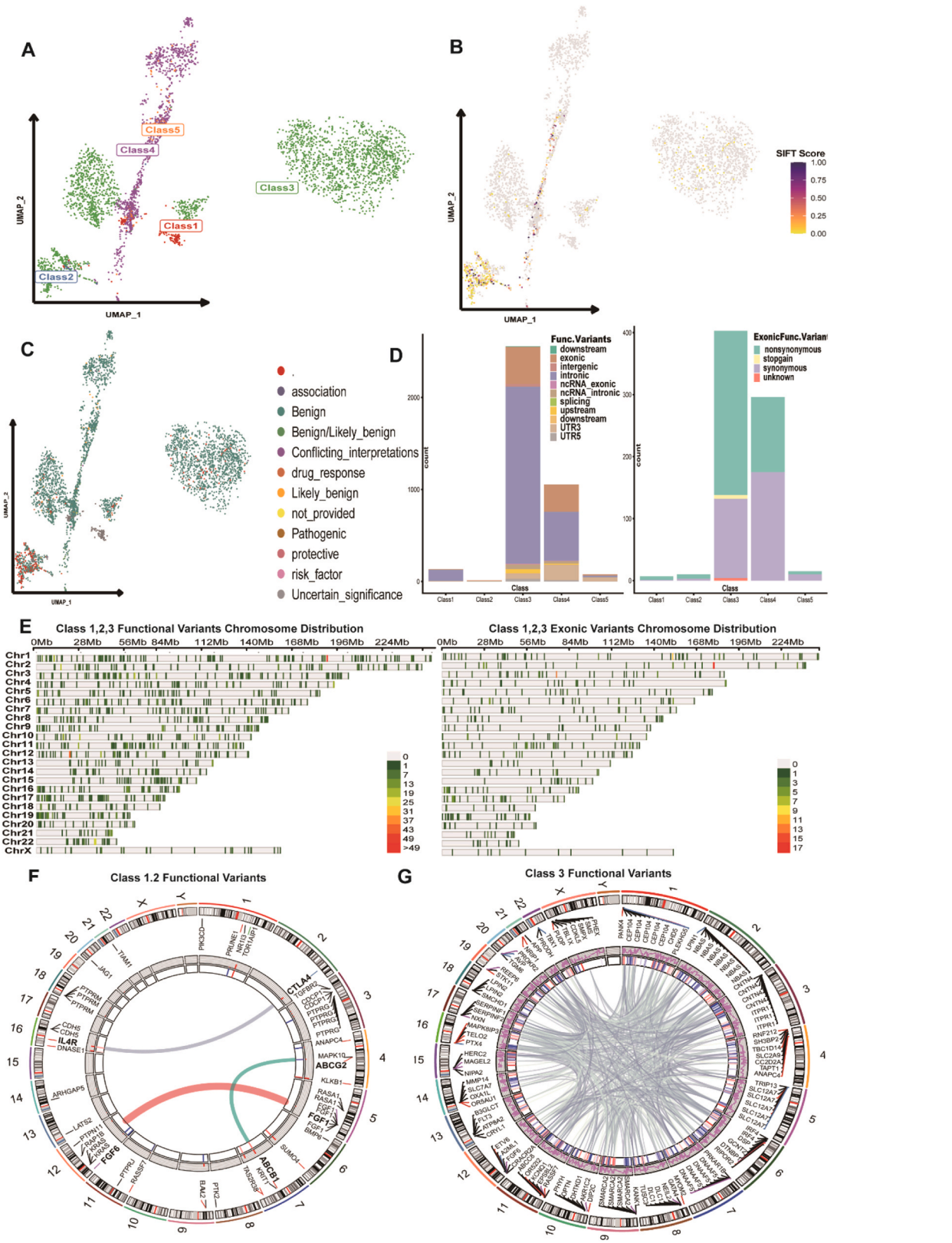

Fig. 2. BRLM mutation classification for a familial CCM WGS. (A) UMAP plot for the BRLM classified CCM SNVs. (B) UMAP of the SNVs with SIFT Scores attached. (C) UMAP for SNVs annotated with Clinvar categories. (D) Functional variants statistics of regulatory and exonic regions for the five classes. (E) Statistics of potential pathogenic variants distribution within functional and exonic regions for Class 1, 2 and 3. (F) Circos plot with low-density functional areas distribution connected by PPI between high CCM risk variants in Class 1, 2. (G) Circos plot with high-density functional areas distribution connected by PPI between uncertain CCM risk variants in Class 3.

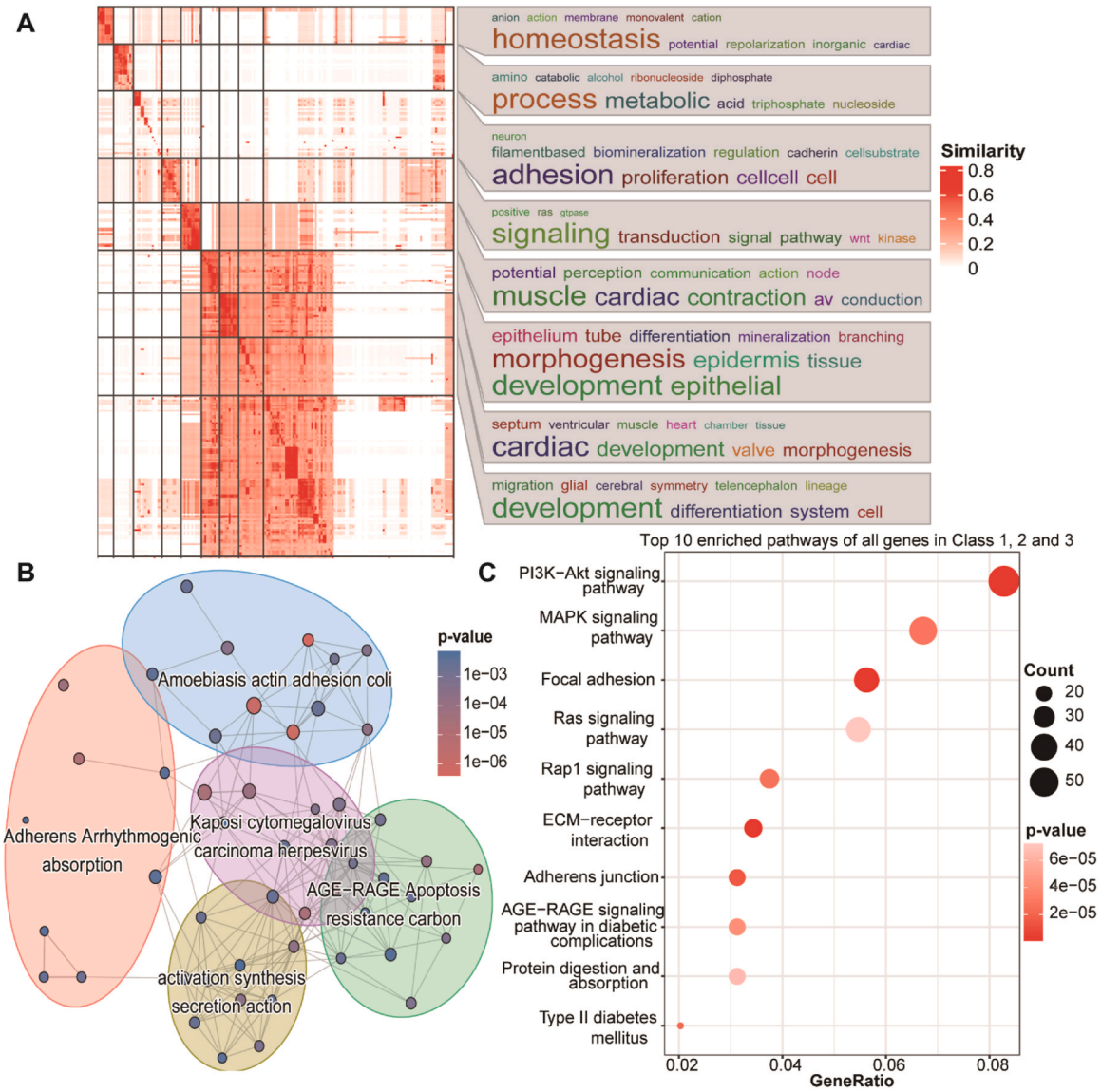

Fig. 3. Enrichment results for mutated genes in Class 1, 2 and 3. (A) Similarity clustering heatmap for enriched pathways with term frequencies exhibited by font size. (B) K-means clustering for the top 50 pathways with the most significant p-values. (C) The top 10 enriched pathways enumeration in terms of p-values.

and proliferation.

In order to further characterize the functional terms, the top 50 pathways were grouped into k-means clusters, illustrated in Fig. 3B. It is notable that the top cluster relevant to cell junction adhesion comprises 17 distinct pathways, while the other three clusters are associated with cGMP-PKG activation, AGE-RAGE signaling, and metabolic correlation. However, these top 50 pathways were still extremely complicated, we thus narrowed them down to feature the top ten pathways with the most significant *p-values*, as highlighted in Fig. 3C.

According to the specific results in Fig. 3C, cell signaling pathways hold a consistently substantial proportion. Particularly, the PIK3-Akt, MAPK, Ras, and Rap1 signaling pathways hold the top four positions, with focal adhesion ranking the third. The aforementioned SNVs identified in Class 1, 2 and 3 exhibited the potential to induce aberrant cell functions, thereby becoming predisposing factors for CCM.

2.4. Simulation of Mutated Genes Perturbation among Pathways

After confirming the involvement of SNVs in enriched pathways, the next study was to address the functional roles of the mutated genes with accumulative effects, which have remained untouched in previous CCM studies. To quantify the functional implications of mutated genes at the pathway level, an algorithm calculated pathway mutations accumulative perturbation score (PMAP score) was adopted [26]. The PMAP score was used to measure the actual perturbation impact on enriched pathways under a candidate gene set encompassing the three classes (Class 1, 2 and 3). A comprehensive list of perturbed pathways along with their PMAP scores for each class is provided in the Supplementary Table S2. According to their PMAP scores comparison, the scores were almost equal between Class 1 and 2, which turned out different from Class 3. We thus take the Class 1, 2 as a whole in the follow-up study regarding perturbation.

Since the gene list in Class 3 contained uncertainties that differ from the well-established lists in Class 1, 2, their proportion of perturbed gene sets were compared within the top 10 highest scored pathways. This result is depicted in Fig. 4A as a tree plot. Based on the pie charts at the end of each tree branch, the perturbation rate of Class 1, 2 is mainly higher or competitive to that of Class 3. Nevertheless, the number of mutated genes in Class 3 far exceed those in Class 1, 2. Moreover, the pathways with remarkably high PMAP scores (including PIK3-Akt, MAPK, Ras, and Rap1 signaling pathways) are also consistent with the top ten enriched results in Fig. 3C. These findings suggest that the SNVs

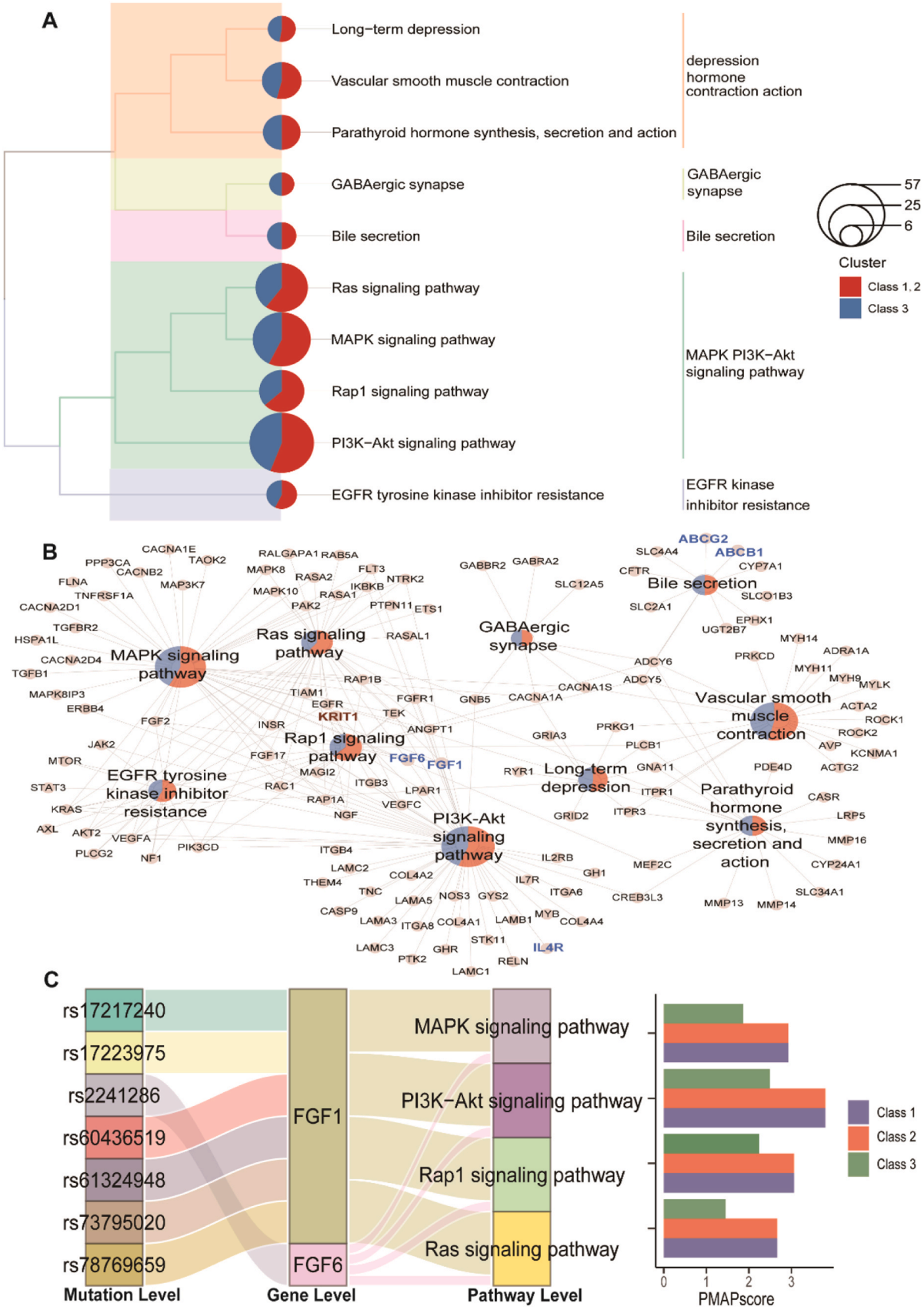

Fig. 4. Pathways perturbation simulation derived from mutated genes in Class 1, 2 and 3. (A) Tree plot for the top 10 pathways with the highest PMAP score, where the pie chart shows the proportion of involved genes from Class 1, 2 and Class 3. It is evident that fewer mutated genes in Class 1, 2 play more important perturbation roles than those in Class 3. (B) Containment relationship for top 10 perturbated pathways and functional domain mutated genes. (C) Sankey plot for risk CCM-related elements in three levels for mutations, genes and pathways.

have an influence on CCM cellular signaling functions, with major effects from Class 1, 2 and minor effects from Class 3.

For primary perturbed genes with functional relevance were identified among Class 1 and 2, we thus constructed a perturbated network focusing on the top ten scored pathways connected by these genes (see Fig. 4B). The ratios of overlap perturbed genes are shown in Supplementary Figure S4, which distinguishes genes that uniquely belong to one pathway or perturb two or more pathways. The five PPI-linked genes (FGF1, FGF6, ABCB1, ABCG2, and IL4R) from Fig. 2F with Class 1 and 2 SNVs are highlighted in bold blue. Furthermore, one of the known pathogenic genes for CCM, KRIT1, is conspicuously present in this network (in brown bold font). However, two intronic variants in KRIT1 were common with frequencies exceeding 0.04 in the 1000 Genomes database as shown in the Supplementary Table S1. Among the pathways, Vascular Smooth Muscle Contraction and GABAergic Synapse exhibit remarkable perturbation scores despite being excluded from the top 10 enriched pathways in Fig. 3C, with *p-values* of 0.002 and 0.03 respectively. This further supports the idea that perturbation algorithms applied in BRLM can help uncover ignored information.

In order to determine the significant risk pathways, the overlap between the top 10 scored (in Fig. 4B) and top 10 enriched (in Fig. 3C) pathways were extracted. Combined with the major effect genes harboring functional SNVs in protein interactions highlighted in Fig. 4B, the elements at three levels (mutation, gene, and pathway level) were established as critical CCM pathogenic routes, elucidated through the three-bucket Sankey graph in Fig. 4C. The first bucket effectively outlines the presence of seven functional SNVs within these genes (six Class 1 SNVs in FGF1 and one Class 2 SNV in FGF6). Notably, all seven SNVs have just been cataloged with rsIDs from NCBI dbSNP, yet none of them bear any reports pertaining to CCM or related biological implications. The second bucket contains a pair of functionally pathogenic mutated genes, namely FGF1 and FGF6 from the fibroblast growth factor family. While the four risk pathways in the third bucket coincide with the top four scored and five enriched results. The overlapping results support the accuracy of both enriched and perturbated pathways, which can improve the integrity of CCM-related mutation knowledge.

To assess the effectiveness of the above CCM-related three-level risk elements, whose classification entities were extracted from PubMed by BioBERT [25]. The PubMed publication proportion statistics from 2010 to 2023 are shown as three trends for each level in Supplementary Figure S5. According to the Supplementary Figure S5A, only SNV rs17217240 had been reported in 2010, while there was no literature support for the others. Regarding the materials for these two genes displayed in Supplementary Figure S5B, over the past two decades, FGF1 has received considerable attention but no relation to CCM, while FGF6 has been scarcely reported. Finally, the statistics for pathway publications reflect overall continuance attention, albeit with varying numerical records as illustrated by the Supplementary Figure S5C. These publication statistics unveil abundant diversity in features for variant classification.

2.5. *FGF1 is the upstream master regulator gene of perturbated pathways*

Based on the enrichment analysis and the perturbation analysis, RNA-seq differential expression geneset between CCMECs and HBMECs introduced before had verified for final major-effect gene determination [44]. FGF1 is found to be highly up-regulated (*p-value*=0 and *log2FoldChange*=1.8), while FGF6 takes no statistical difference (*p-value*=0.71 and *log2FoldChange*=−0.6). The detailed differentially expressed results for genes engaged in perturbated pathways are illustrated in the Supplementary Table S3. Herein, we dive into the details of FGF1 from multi-omics as shown in Fig. 5, including the scRNA-seq expression clusters (Fig. 5A), the WGS mutant transcripts (Fig. 5B), the RNA-seq expression profiles (Fig. 5C), and the perturbated pathways (Fig. 5D).

For exploring expressed cell groups of FGF1, scRNA-seq was carried out on a pair group of CCM mouse model under two normal conditions and two deletions of Pdcd10 [36]. The UMAP plots for joint clustering are displayed in Fig. 5A, with 16,220 cells from the two Pdcd10-wt mice and 19,135 cells from the two Pdcd10-ko mice after low-quality cells filtering. Based on gene-marked UMAPs among 12 cell clusters, FGF1 identifies special expression in cluster 8 whose marker genes (Aqp4, Gfap, and Slc1a2 colored in blue) are primarily associated with astrocytes. This cluster also includes the receptor genes for the FGFR family, with Fgfr3 and Fgfr1 (colored in blue) exhibiting specific and significant expression. Additionally, Atp1b1, a marker for capillaries (colored in red) is also highly expressed in this cluster. Recent research has indicated that astrocytes propel neurovascular dysfunction during cerebral cavernous malformation lesion formation [29].

All six SNVs in FGF1 were classified as Class 1, corresponding to two transcripts: FGF1–008 (transcript ID: ENST00000378046) and FGF1–013 (transcript ID: ENST00000411960). Both transcripts performed protein-coding functions with most mutants situated on the longer one (FGF1–013), see transcripts distribution in Fig. 5B. All the transcripts for FGF1 are listed in the Supplementary Figure S6 referenced from GENECODE19 (https://www. gencodegenes.org/human/release_19.html).

To further explore the cumulative influence between the major-effect gene FGF1 and other minor-effect genes found by BRLM, differential expression profiles whose parameters taken from the Supplementary Table S3 are constructed in Fig. 5C. The minor-effect genes were detected in positive sites across Class 1, 2 and 3, along with the corresponding perturbated pathways from Fig. 4. Most of these genes show significant differential expression in both $p\text{-}value < 0.05$ and $|log2FoldChange| > 1$ indicated by read dots in Fig. 5C.

According to the categorized multi-effect genes based on mutation and expression, an expanded biological regulatory landscape featuring partially function exploration is conducted in Fig. 5D. FGF1, highlighted in orange, demonstrates a master regulatory role, while the remaining genes are associated with minor impact SNVs across Class 1, 2 and 3. As we can see from Fig. 5D, FGF1 is located upstream, indicating the direct perturbation of the four signaling pathways denoted by orange boxes. Importantly, among them the downstream of Rap1 Signaling pathway is positioned to one of the known CCM pathogenic genes, KRIT1. These four signaling pathways play critical roles in cell growth and tissue development, which can impact vascular integrity.

In addition to the direct action of FGF1 on the four signaling pathways, two indirect processes, namely Vascular Smooth Muscle Contraction and GABAergic Synapse indicated by green boxes, are regulated by $Ca2+$ levels via Calcium Signaling Pathway, a downstream fundamental cellular signaling process of FGF1. Above findings indicate that FGF1 may activate downstream signals, which suggests a significant role in the occurrence and development of familial CCMs. In order to improve the reliability of our model, we have provided further evidence for the regulatory function of FGF1 in CCM. The supporting literatures are chronologically rearranged in Figure S6 and detailed in Section S2 of the Supplementary file. These materials provide a comprehensive overview of FGF1, emphasizing its potent role in inducing angiogenesis in the brain by regulating multiple growth factor signaling pathways. The development of FGF1 as a powerful stimulator of angiogenesis for various brain-related diseases will be emphasized in the Discussion section.

According to the previous studies, the technologies of genetic or chemical CCM induced models are still immature. The typical cell experiment was to cultivate cells (mostly endothelial cells) from CCM patients for further research based on the original RNAseq results in Fig. 5C [44]. On the other hand, there are diverse findings showing FGF1's pivotal role in angiogenesis, as detailed in Section S3 of the Supplementary file. The main results and conclusions are summarized as follows. Firstly, FGF1 is found to induce the proliferation and migration of endothelial cells at the cellular level [61], whose results are displayed in Supplementary Figure S7. Secondly, the FGF1 morpholino

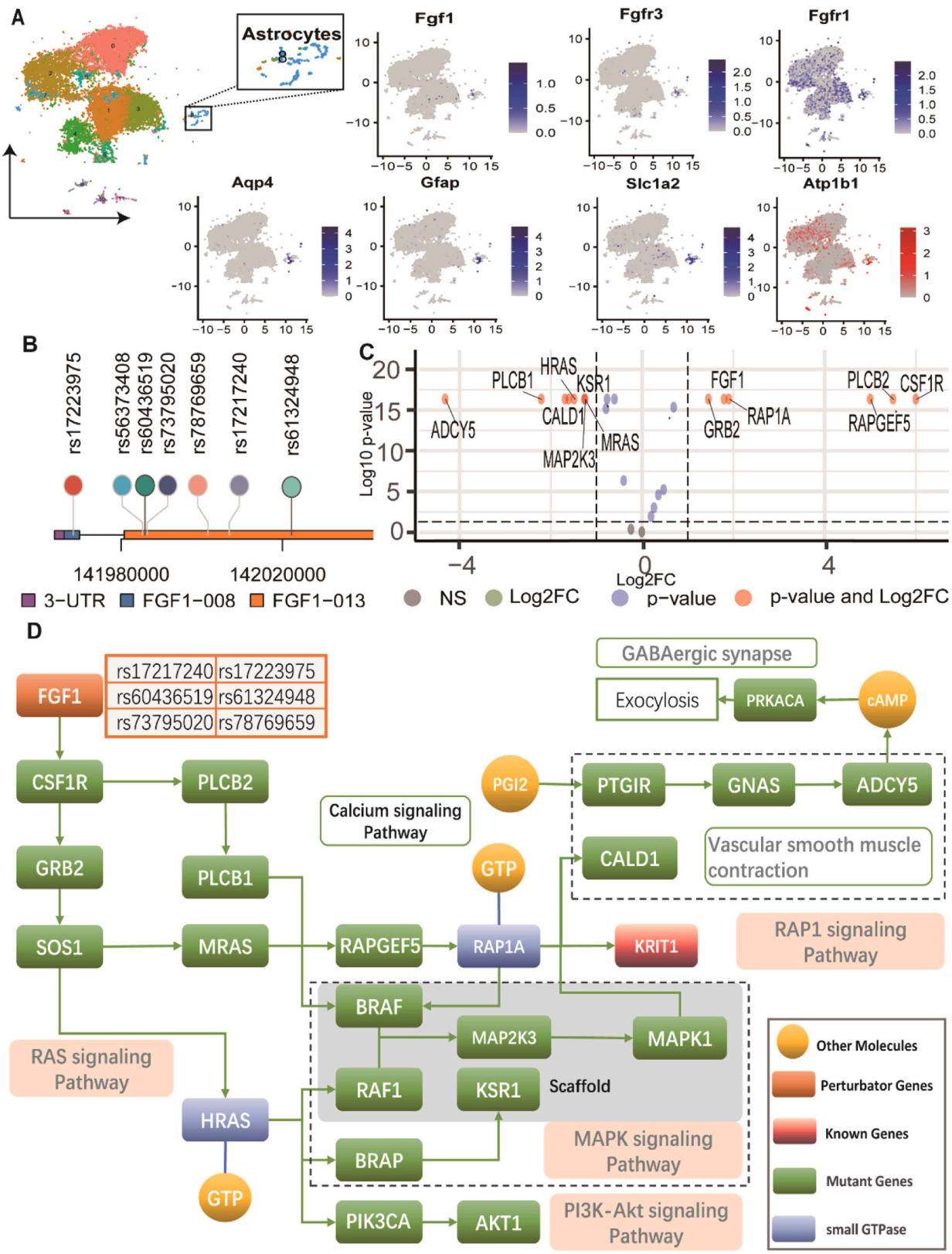

Fig. 5. The FGF1 acts as the master regulator gene upstream of perturbed pathways from multi-omics results integration. (A) FGF1 specific expressed in Astrocytes Cluster from scRNA-seq. The markers for "Astrocytes" cluster are colored in blue, while the marker for "Capillaries" is in red. (B) FGF1 mutation sites distribution in two mutant transcripts from WGS. (C) Differential expression profiles for multi-effect genes from RNA-seq. (D) Main effect gene FGF1 with multiple functional variants located upstream of peaked genes from perturbated and enriched pathways, reacted with mutated genes in Class 1, 2 and 3 including KRIT1, one of the three known CCM pathogenic genes. Multi-connection mutated genes in the same pathway are outlined with dashed lines.

knockdown zebrafish embryos display anomalous cell aggregation in the intermediate cell mass [46], whose results are shown in Supplementary Figure S8. Last but not least, in a rat model involving the implantation of collagen-suspended beads into exposed femoral pedicles, a notable enhancement in vascular density is observed at 1 and 6 weeks compared to the bolus administration of FGF1 [32], whose results are presented in Supplementary Figure S9 and Supplementary Figure S10.

3. Materials and Methods

The BRLM web server is available at http://1.117.230.196 and the source codes are available at https://github.com/wangyiqi806643897/BRLM. The raw sequence data reported in this paper have been deposited in the Genome Sequence Archive (Genomics, Proteomics & Bioinformatics 2021) in National Genomics Data Center (Nucleic Acids Res 2022), China National Center for Bioinformation / Beijing Institute of Genomics, Chinese Academy of Sciences (GSA-Human: HRA005489) that are publicly accessible at https://ngdc.cncb.ac.cn/gsa-human. The other results are shown in Supplementary files.

3.1. CCM family samples collection

The samples for these eight family members were collected at Shiyan Taihe Hospital, Hubei, China, whose family genetic pedigree is presented in Supplementary Figure S2A. Informed consent was obtained from all subjects, and this study was approved by Taihe Hospital. The inclusion and exclusion criteria for the data are based on the MRI results as shown in Supplementary Figure S2B-I, with four individuals in the case group and four in the control group. Each MRI image was reviewed by three independent doctors who produced the official reports and the cross-referenced slices. The case group comprises the proband in Supplementary Figure S2B, a 34-year-old male patient with CCM, and three affected first-degree relatives, including his mother (died of CCM) in Supplementary Figure S2C, aunt (52 years old) in Supplementary Figure S2D, and brother (30 years old) in Supplementary Figure S2E. The control group comprised four relatives from his cousin's family in Supplementary Figure S2F-I. Seven samples were obtained from collateral blood, except for the mother, who unfortunately passed away due to the condition. Her sample was obtained from a block of angioma tissue.

3.2. Whole genome sequencing

Genomic DNA extracted from blood was assessed for quality using PicoGreen and gel electrophoresis. At least 10 μg of non-degraded DNA was provided for WGS. Tissue extraction used the Maxwell 16 Tissue DNA Purification Kit (Promega), which followed the manufacturer's instructions and utilized 10 mg of tissue. Additional quality controls were conducted, including assessing DNA purity and integrity through agarose gel electrophoresis. Furthermore, DNA purity was determined using Nanodrop detection (OD 260/280 ratio), and DNA quantification was carried out with Qubit 2.0. To shear approximately 300 ng of high-quality DNA samples (OD 260/280 =1.8–2.0), a Covaris S220 Sonicator (Covaris) was used to generate fragments of ~350 bp. The fragmented DNA was purified using Illumina's Sample Purification Beads. Adapter-ligated libraries were prepared using TruSeq Nano DNA Sample Prep Kits (Illumina) in accordance with the Illumina protocol. The sequencing was conducted on an Illumina HiSeq system for 2 * 150 paired-end sequencing at Novogene in Wuhan, China.

3.3. CCM family WGS analysis

After quality control and trimming by Fastp [9], the dataset consisting of high-quality clean sequences in fastq format was obtained. Subsequently, sequence alignment to the GRCH37 reference genome was executed using SAMtools with default parameters, achieving a mapping rate of over 90%. The Germline short variant discovery pipeline of GATK version 4 (GATK4) [5] was then employed.

To evaluate the feasibility of each site in our data and identify false positives, VQSR of GATK was employed in two modes: SNP mode and Indel mode. For each mode, distinct reference databases were assigned to the corresponding argument sets, and their parameters are detailed in Table 2. In SNP mode, the utilized databases included HapMap3.3 [10], OMNI2.5 [40], 1000 Genomes [12], and dbSNP (http://www.ncbi.nlm.nih.gov/projects/SNP/). While in Indel mode, the databases consisted of dbSNP (http://www.ncbi.nlm.nih.gov/projects/SNP/) and Mills Gold [30]. Each of the aforementioned databases necessitates the determination of four key parameters: "known" (whether the data is used as known variation for marking), "training" (whether the data is used for training), "truth" (whether the data is used as the ground truth for verifying), and "prior" (the weight of the data set in model training, or prior likelihood).

3.4. SNVs annotation texts preparation

To annotate SNVs for model training, a multi-step pipeline was employed. Initially, this pipeline established a comprehensive annotation process involved three SnpEff and seven ANNOVAR referencing databases. Simultaneously, relevant gene information was curated from prominent genetic databases, encompassing details such as function, expression, phenotype, and pathway.

The annotation databases encompassed resources like Clinvar [24] (version 20220320), 1000 Genomes (version 1000g2015aug_ALL and 1000g2015aug_AFR) [12], dbSNP (http://www.ncbi.nlm.nih.gov/projects/SNP/), SIFT [28], GWAS [53], Kaviar [15], eigen [22], and gnomAD [23].

Besides the above database annotations, further gene information had also added to the descriptions of the SNVs accordingly, including the functions, expressions, distributions, phenotypes presented in GeneCards (https://www.genecards.org/) and NCBI (https://www.ncbi.nlm.nih.gov/), and the pathway terms presented in Gene Ontology (GO, http://geneontology.org/) and Kyoto Encyclopedia of Genes and Genomes (KEGG, https://www.genome.jp/kegg/).

All these genetic insights were incorporated into natural language descriptions for BioBERT with the input format demonstrated in Table 3 and an example shown in Supplementary Table S5. The input files are available for downloading from the web server (http://1.117.230.196/results_download), along with the source codes (https://github.com/wangyiqi806643897/BRLM).

3.5. TCGA verified dataset

Initial verification of SNVs by BRLM was conducted using data

Table 2
Data parameters of GATK VQSR.

Resource	"known"	"training"	"truth"	"prior"	Reason
SNP mode: HapMap	false	true	true	15.0	Strict quality control and experimental verification
OMNI	false	true	true	12.0	Gold standard for genotypes
1000 G	false	true	false	10.0	Deficiency of comprehensive experimental verification
dbSNP	true	false	false	2.0	Submitted results without rigorously verification
Indel mode: Mills Gold	true	true	true	12.0	Verified dataset
dbsnp	true	false	false	2.0	Same as SNP mode

Table 3
BioBERT input text format.

	Anno Category	Text Format
Annotation	Pos	Chr:{text} Start:{text} End:{text} Ref:{text} Alt:{text}
	Ref	Func.refGene:{text}Gene.refGene:{text}GeneDetail. refGene:{text} ExonicFunc.refGene:{text} AAChange. refGene:{text}
	Database	CLNALLELEID:{text} CLNDN:{text} CLNDISDB:{text} CLNREVSTAT:{text} CLNSIG:{text} 1000g2015aug_ALL:{text} 1000g2015aug_AFR:{text} avsnp150:{text} avsift:{text} GWAVA_region_score: {text} GWAVA_tss_score:{text} GWAVA_unmatched_score:{text} Kaviar_AF:{text} Kaviar_AC:{text} Kaviar_AN:{text} gnomAD_exome_ALL:{text} gnomAD_exome_AFR: {text} gnomAD_exome_AMR:{text} gnomAD_exome_ASJ:{text} gnomAD_exome_EAS: {text} gnomAD_exome_FIN:{text} gnomAD_exome_NFE:{text} gno-mAD_exome_OTH: {text} gnomAD_exome_SAS:{text}
	Info	NCBI_Summary:{text} GeneCards_Summary:{text} Swiss-Prot_Summary:{text} GO:{text} KEGG:{text}

obtained from twelve cancers within the Cancer Genome Atlas (TCGA). Cancer types from TCGA, accompanied by their abbreviations, included Adrenocortical carcinoma (ACC), Breast invasive carcinoma (BRCA), Bladder urothelial carcinoma (BLCA), Colon adenocarcinoma (COAD), Cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC), Cholangiocarcinoma (CHOL), Lymphoid neoplasm diffuse large B-cell lymphoma (DLBC), Esophageal carcinoma (ESCA), Glioblastoma multiforme (GBM), Kidney chromophobe (KICH), Pan-kidney cohort (KICH+KIRC+KIRP) abbreviated as (KIPAN), and Brain lower grade glioma (LGG). Following verification of the aforementioned cancers, mutation annotation format (MAF) files were downloaded, with each line representing a single variant. These original MAF files were then converted into avinput format by convert2annovar.pl in ANNOVAR (http://annovar.openbioinformatics. org/en/latest/).

3.6. Analyzing environment configuration

WGS analysis was conducted on a server having 128 G RAM and two Silver 4114 CPUs (40 cores in total), installed with CentOS 7.0. SAMtools mapping was performed using 8 cores, while the SNVs calling step utilized 16 cores.

BRLM was implemented by using python3.8 with the deep learning framework of PyTorch1.9. The learning rate was optimized by Adam with an initial value of 10^{-2}, a reduction factor of 0.1, and a batch size of 8. The model was evaluated on a cluster having four NVIDIA V100 GPUs.

3.7. BRLM's workflow

The overall workflow of BRLM is shown in Fig. 6. After obtaining annotation texts for SNVs, the BioBERT encoder and ResNet classifier pipelines were developed for classification, with the training on TCGA and testing on CCM datasets.

3.7.1. Encoding variants

The encoder for SNVs was achieved by BioBERT which stands out as an extensively employed model in the field of natural language processing for medical and biological purposes. It was developed as a specialized iteration past the BERT (Bidirectional Encoder Representations from Transformers), initially pretrained on English Wikipedia and BooksCorpus [13].

Nevertheless, owing to the significant presence of biomedical-specific proper nouns (such as KRIT1, c.369 A>G) or terms (like exonic, intronic), BERT's general models struggled to achieve a satisfactory performance. This limitation prompted the emergence of BioBERT, tailored for tasks involving biomedical text mining. This variant was pretrained on PubMed abstracts (available at PubMed: https://pubmed.ncbi.nlm.nih.gov/) and full-text articles from PubMed Central (accessible at PMC: https://www.ncbi.nlm.nih.gov/pmc/). Considering that descriptions of SNVs for annotation could also be categorized as domain-specific natural language expressions, we leveraged BioBERT's entity extraction capabilities to supplant the manual literature querying process in the interpretation of SNVs.

All variants were annotated within sentences, activating the encoding module for deep learning purposes. Regarding BRLM, we utilized BioBERT to encode annotated texts, yielding 728-dimensional vectors per SNV. The employed model version was biobert-base-cased-v1.1, with a batch size of 100 and the "mean" pooling algorithm for pooler output. Regarding pooling, it is used to reduce the size of the feature maps and avoid sacrificing too much information [6]. The "mean" pooling focuses on overall features with less influence from outliers, making it more robust than the "max" pooling operation [20].

Moreover, instructions were incorporated into each annotated sentence subsequent to the annotations generated by ANNOVAR. The

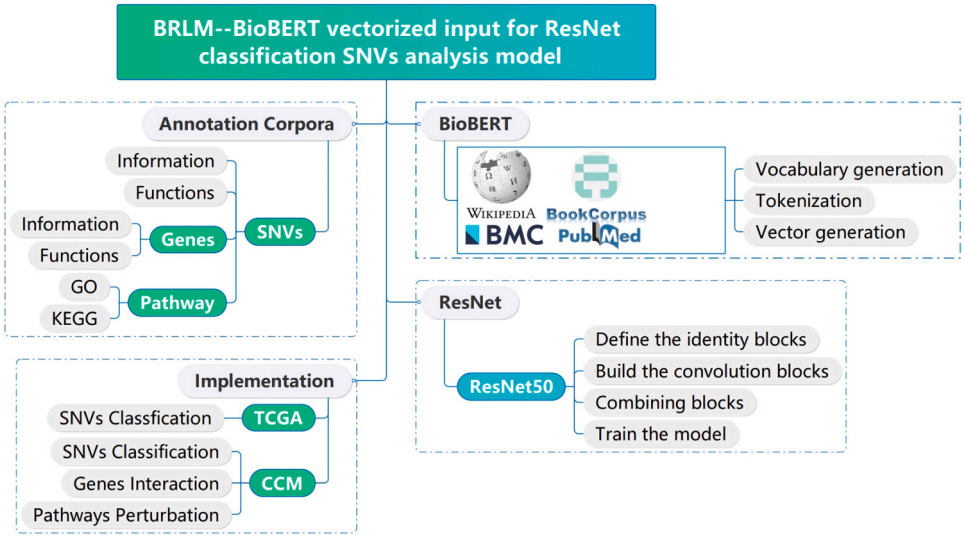

Fig. 6. BRLM workflow for variant annotations classifying. Starting with annotated data wrangling, embedded vectors are constructed by BioBERT, which are classified by ResNet50 for distinct datasets.

instruction format followed the pattern of "Disease name" + "mutation sites". These disease names were contingent upon the data sources, either TCGA cancer species or CCM.

3.7.2. Classifying embedded vectors

The ResNet50 [51] was borrowed to identify pathogenic mutations, whose architecture was illustrated in Fig. 1A. The neural network depth designed for mutation site analysis significantly outperformed traditional neural networks in handling SNVs' large data inputs, preventing gradient vanishing and performance degradation.

To elucidate the unique structure of ResNet, we conducted a comparative summary of its specific advantages against two other classical convolutional neural networks (VGG and Inception), as depicted in Fig. 7A. The skip connection proposed in ResNet is a significant breakthrough of deep neural network, which provides a shortcut for gradients to flow more easily through the network during back-propagation. Instead of passing through every layer, gradients can take a shortcut and directly propagate to deeper layers. This helps to mitigate the vanishing gradient problem, allowing the network to learn more effectively even as it becomes very deep.

In BRLM, the classification architecture was similar to ResNet50 but differed in network size. Precisely, we tailored our network input to accommodate the 728-dimensional vectors generated from BioBERT, and adjusted the subsequent layers accordingly. The detailed structures are outlined in Fig. 7B, recording the kernel size, quantity of kernels, and stride size in each layer of the four residual blocks. In regards to the stacked residual blocks, each block comprised of three convolutional layers with a "skip connection" that bypassed the three layers within each individual block. The goal was to achieve a five-class classification, as demonstrated in Fig. 1C. For this unbalanced dataset, the class weight was applied in the "CrossEntropyLoss" function, resulting in the loss being calculated as

$$\mathscr{L} = \sum_{i} \frac{N_i}{N} \sum_{j} y_j^i \log \quad p_j^i \tag{1}$$

where N_i is the number of samples in class i with $N = \sum N_i$, y_j^i is the label of sample j in class i, and p_j^i is the predicted probability of sample j in class i.

The training, validation, and test sets were divided using a 7:1:2 splitting ratio. This ratio had been considered to be the optimal criteria [33]. The evaluations were performed based on a pre-labeled pan-cancer dataset sourced from TCGA by accuracy, precision, recall, mean average precision (mAP), and F1-score. Precisely, the predictions are recognized in the following manner: (i) True positive (TP) is the number of SNVs classified as pathogenic variants correctly; (ii) False positive (FP) is the number of SNVs classified as pathogenic variants incorrectly (unrelated mutation in fact); (iii) False negative (FN) is the number of SNVs deemed as non-pathogenic variants incorrectly; and (iv) True negative (TN) is the number of SNVs deemed as non-pathogenic variants correctly. The

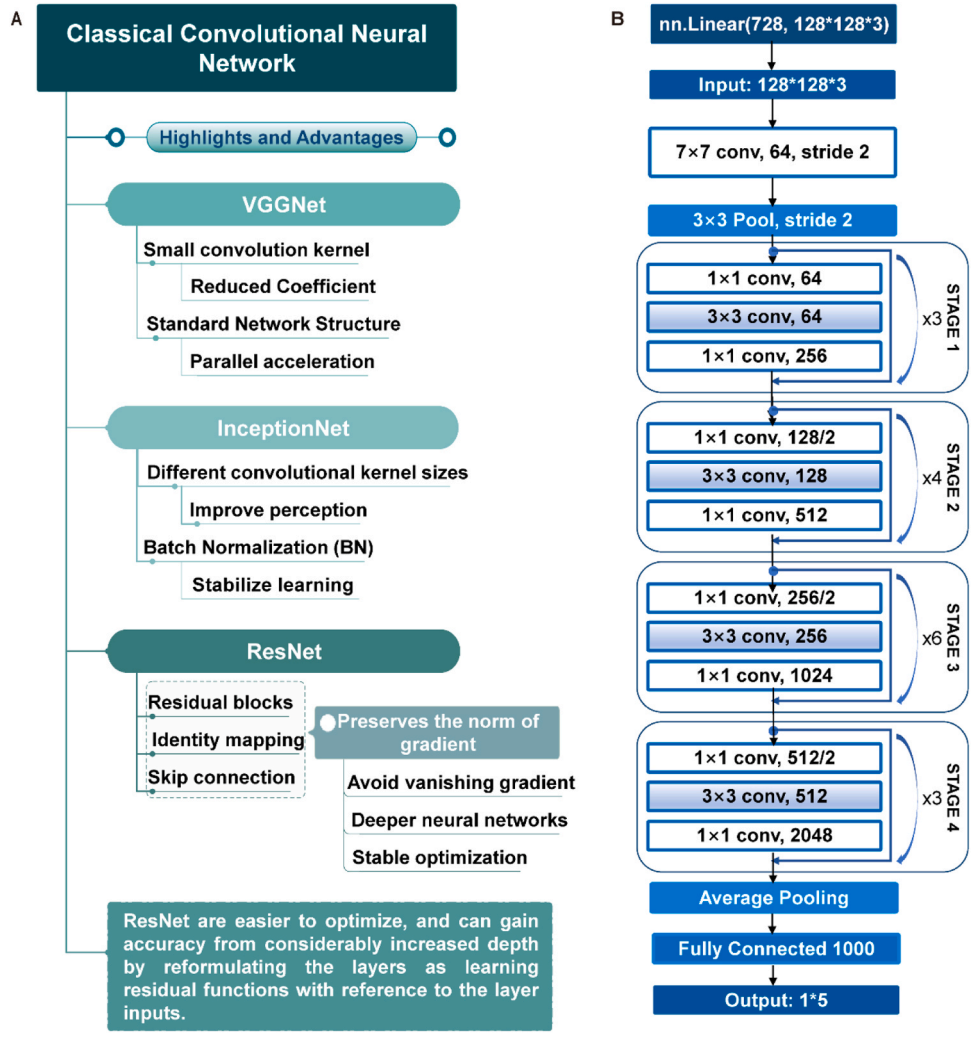

Fig. 7. Particular structure of ResNet compared with classical convolutional neural networks and ResNet50 architecture diagram in BRLM. (A) ResNet residual network with skip connection can solve gradient vanishing problem. (B) ResNet-50 architecture constructed in BRLM for SNVs classification.

formulas are as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{4}$$

$$\text{F1} - \text{score} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5}$$

Emphasizing the primacy of training BRLM from TCGA, we eventually incorporated CCM variant vectors for testing.

3.8. Visualization of Classification Results

The basic plots in this paper were created in R, using the ggplot2 package's built-in functions. This package was used for basic graphing methods, such as barplots, dotplots, roseplots, and Sankey diagrams. Furthermore, different types of graphs were developed employing certain packages, which will be introduced in later sections.

3.9. UMAP construction for variants

Utilizing BRLM input vectors extracted from BioBERT, it was observed that the variants were present in a 728-dimensional space, which makes retaining the global structural information derived from ResNet50 classification results. To effectively map the entire set of variants onto a two-dimensional space, the Uniform Manifold Approximation and Projection (UMAP) dimensionality reduction technique was employed.

Variants embedding vectors from TCGA or CCM datasets can be utilized as input entities for the "umap" package in R4.1. All UMAP graphs presented in this paper were generated using 5 components representing five distinct classes and 20 neighbors, while considering multiple categorical labels including class, SIFT score, Clinvar significance, and density.

3.10. Variety of genetic mutations statistics

The statistical charts for SNVs were generated using the R4.1 package known as "maftools", which is designed to process MAF files. The "bcftools" was used to convert ANNOVAR multi-anno files from VCF to MAF. Finally, "plotmafSummary" function was employed for charts generating.

3.11. SNVs chromosome distribution statistics

Two types of chromosome distribution charts were utilized: the first employed a chromosome distribution strip map using the CMplot package in R4.2, while the second utilized a circos plot.

The circos chromosome plot was generated with the RCircos package for R version 4.2 and its chromosome track was constructed based on the UCSC HG19 Human CytoBandIdeogram data, which was imported along with the corresponding gene set. Additional tracks were pre-built, each with their own subfunction calls. The middle two frequency statistics channels presented the 1000 Genomes and genomAD records for each mutation site and converted them into input invocation matrices beforehand. Internal lines were aligned with PPI pairs from the STRING database.

3.12. Genes enrichment

The ClusterProfiler package was primarily utilized for enriching mutant genes in the R4.2 software package with the "dotplot" function of Fig. 3C visualizing the top 10 pathways, and the "treeplot" function was called to produce tree diagram in Fig. 4A. In order to summarize entire significant enriched results, the simplifyEnrichment package was used to cluster similarity matrices calculated by "term_similarity" function with a new method named "binary cut" [16]. With the similarity matrix, we can directly apply "simplifyEnrichment" function to perform partition around medoids (PAM) with two groups on the similarity matrix in each iteration step. The similarity clustering algorithm was applied to cluster pathways, as depicted in the heatmap of Fig. 3A.

Similarly, the K-means clustering algorithm utilized the "emapplot_cluster" function from the ClusterProfiler package to create the network displayed in Fig. 3B.

3.13. Genes perturbation algorithm

In order to measure the pathway mutations perturbation levels, the cumulative effect of mutated genes were quantified by PMAP scores using an R package known as PMAPscore (https://cran.r-project.org/web/packages/PMAPscore/vignettes/PMAPscore.html). The package's "get_mut_status" function has been modified to accommodate variants in Class 1, 2 and 3. Subsequently, the cumulative effect of genetic mutations on pathways is utilized to determine the PMAP scores. This score employs a standard cumulative perturbation measurement to capture the positioning and impact of genetic mutations on pathways. The formulas of perturbation scores for genes and their cumulative effect in pathway are as follows.

Gene's Perturbation score:

$$\text{GMPscore}(g_i) = 1_i(g_j) + \sum_{j=1}^{N} \beta_{ij} \frac{\text{GMPscore}(g_j)}{N_{ds}(g_j)} \tag{6}$$

where $\text{GMPscore}(g_i)$ denotes the perturbation score of mutated gene g_j, $1_i(g_j)$ is an indicator function with $1_i(g_j) = 1$ if g_j belongs to class i, otherwise 0, β_{ij} is the relationship between genes g_i and g_j (if g_j is directly interacted with g_i, $\beta_{ij} = 1$, else is 0), N is the total number of genes in pathway p_i, $N_{ds}(g_j)$ is the number of genes at the downstream of gene g_j.

Pathway's Perturbation score:

$$\text{PMAPscore}(p_i) = \frac{\sum_{k=1}^{N} \text{GMPscore}(g_k)}{N_{dc}} \tag{7}$$

where $\text{PMAPscore}(p_i)$ denotes the perturbation score of pathway p_i, N is the total number of genes in pathway p_i, $\text{GMPscore}(g_k)$ denotes the perturbation score of gene g_k in the pathway p_i, N_{dc} denotes the number of genes that have mutated in the pathway p_i.

3.14. Public scRNA-seq data analysis

The public scRNA-seq raw data were obtained from the Gene Expression Omnibus database with ID GSE155788 (https://www.ncbi.nlm.nih.gov/geo/). The Data analysis relied on Seurat (version 4.0.5), the filtering criteria was nFeature_RNA < 300 or > 12000 with $> 15\%$ expression of mitochondrial genes. After completing quality control, the next step was normalization. This was accomplished by using the "LogNormalize" method with default scale factor and log-transformation. Subsequently, the "FindVariableFeatures" algorithm calculated a subset of features that yielded significant cell-to-cell variation in the dataset. This function returned 2000 default features per dataset, which will be used in downstream analysis.

Prior to the dimensionality reduction, a linear transformation (called scaling) was applied with the "ScaleData" function. Next, the "RunPCA" function was invoked for linear dimensionality reduction on the scaled data. In order to determine the dimensionality, the JackStraw procedure was executed for principal components selection. As a result, the remaining cells were clustered together with the npcs of 30 and a

resolution of 0.3. The final clusters were annotated based on the markers shown in [36].

4. Discussion

We have shown that BRLM, constructed from BioBERT encoder and ResNet classifier, can serve as a SNVs annotation learning model to assist variants classification and interpretation. BRLM was designed with the aim of conducting mutation sites analysis associated with various diseases. The accuracy of BRLM was verified on twelve cancer types in TCGA, whose progressive accuracy in each epoch was observed across all cancer datasets. The generalizability was validated on CCM sequencing analysis, with SIFT scoring and Clinvar annotation validated on the classification results. Following the perturbation scoring algorithm to quantify the multi-effects of mutated genes in pathways, an upstream master regulator gene FGF1 was found supported by the astrocyte cluster of scRNAseq and differential expression of RNAseq. Our results demonstrate the feasibility of BRLM in classifying SNVs and contribute to the discovery of major pathogenic factors.

BRLM innovatively employed the NLP algorithm in WGS pathogenic information mining, making the image classifier applicable to variants classification. The application of NLP models in single-cell sequencing datasets has gained popularity, but this technique is still limited in clinical laboratories. Some other studies have focused on molecular information mining from published biological texts. It should be NLP's first use in genome sequencing to solve clinical problems. As for this CCM family who visited for procreation guidance, we had to categorize all SNVs with no reasonable pathogenicity found in known CCM-caused genes, and further employed the perturbation scoring algorithms to quantify the accumulation effects of multi-class mutations in pathways. Finally, the pathogenic mechanisms in CCM were elucidated, showing that a major gene FGF1 could contribute to CCM development alongside the minor genes' effects. This suggests that the applicability of this model can be widely used in clinical genetic counseling.

FGF1 can be detected in recent sequencing results owing to the broad usage of bulk sequencing, other than previous CCM omics studies that have only probed three known genes. For instance, the direct PCR implementation revealed five variants in the CCM3/SERPINI1 asymmetric bidirectional promoter [41], and the analysis of coding exons identified a novel missense mutation, c .422 T > G, in CCM3 [42]. With the widespread adoption of next-generation sequencing (NGS), an expanded set of genes can now be examined. When the mini-bulk RNA sequencing data of MAP3K3 mutant individuals were compared with those of MAP3K3 WT individuals during fCCM3 lesion formation, FGF1 was found to be another up-regulated gene, indicating the activation of ERK1 and ERK2 cascades [56]. Additionally, the downstream of FGF1, namely TGFBR2 and ACTG2, were found to be consistent with our perturbed results, indicating shared pathways such as Hippo and MAPK signaling [43]. Subsequently, Fgf1 expression decreased after GJA1–20k altered the endothelial cell transcriptome with hypermethylation of its gene body in animal experiments [49].

The FGF family, comprising six subfamilies (FGF1, FGF4, FGF7, FGF8, FGF9, and FGF19) [35], plays a critical role in embryonic development and organogenesis by maintaining progenitor cells and promoting their growth, differentiation, survival, and patterning [21]. The FGF1 subfamily, inclusive of FGF1 and FGF2, serves as potent angiogenic inducers that control multiple growth factor signaling. The FGF1 subfamily regulates vessel formation [34], promotes strong angiogenic responses [27], and induces vessel maturation [14]. Notably, it has been implicated as a potential instigator of aberrant angiogenesis [18], which could be linked to the pathogenesis of arteriovenous malformations [19] and other cerebral vascular anomalies [45]. As the only gene in the FGF family undergoing clinical trials [4], FGF1 shows promise for stimulating blood vessel growth in the brain and addressing various brain-related diseases, including intracranial aneurysm [60], Alzheimer's disease [52], brain tumors [3], brain injuries [4], and ischemic stroke [63].

According to the multi-omics data of FGF1 in CCM and its relative biological functions, we speculated that the BRLM analyzed results of FGF1 was an upstream regulatory gene with clinical implications in CCM. The CCM-related elements presentation was a landscape of perturbed pathways with FGF1 positioned upstream and interacting with downstream mutated genes, including a known CCM pathogenic gene, KRIT1.

Moreover, BRLM is just constructed from genomic data, an integrating model can be anticipated with multimodality data (such as clinical records, images, vital signs monitoring, etc.), which may be expected to provide more clarity in understanding pathogenicity mechanisms and elucidating functional genes. Our next efforts will be focused on refining information consolidation and multimodal learning exploitation. Alternatively, directions of future work on BRLM will encompass three models. Firstly, the initial text-coding module will be upgraded into a time-dependent version so that the progression of diseases can be accommodated [1]. Secondly, an image encoding module will be added to cope with visual data [37]. Finally, a transformer is conceived so that various modalities, such as texts and images, can be handled simultaneously [2]. Our intention is to construct a medical knowledge graph illustrating the topological relationships among clinical symptoms, laboratory indicators, biological markers, related diseases, and targeted drugs [57]. The prototype functions have been integrated into the public interface, incorporating targeted drug retrieval and ceRNA network construction. Further development of these functionalities will be presented in our following studies.

In conclusion, this study offers a novel insight for pathogenic factors exploration through biomedical language.

analysis, potentially assisting manual retrieval methods and making up complicated biological experiments.

5. Conclusions

In this study, a biomedical language learning model called BRLM was developed to classify and link SNVs, with the goal of assisting in the labor-intensive tasks of interpretation and integration. The pipeline was utilized to classify variants into five categories, as defined by ACMG guidelines with multi-class interaction network construction. Comprehensive databases were compiled containing annotations that describe variants in biomedical natural language. To facilitate this, we employed the BioBERT, which primarily focuses on entity recognition during embedding. Then the encoded vectors were used to train a convolutional neural network. This model was trained on 12 TCGA datasets and tested on a familial CCM dataset. From the classified results, we performed pathway variants accumulative perturbation analysis, and found a master regulatory gene, FGF1, that could be highly related to CCM. The core contribution of this study resides in the integration of a large language model into a variant classifier, where the former is able to capture the essential information from the massive input data while the latter guarantees better usage of the acquired information. The effectiveness of this protocol has been verified by multi-omics sequencing analysis. Moreover, a web server has been realized to facilitate the broad usage of the proposed model. This study highlights the potential of our approach to comprehend the complex interplay of genetic variants within biological terms.

Institutional Review

The study was approved by the Institutional Review Board (or Ethics Committee) of Taihe Hospital, Shiyan, Hubei, China (protocol code 2023KS33, September 15th, 2023).

Funding

This research was funded by the National Natural Science

Foundation of China (32070973 and 32060150).

CRediT authorship contribution statement

Yiqi Wang: Conceptualization, Methodology, Software. **Jinmei Zuo:** Data collection and sampling. **Chao Duan:** Data curation. **Hao Peng:** Data collection and analysis. **Jia Huang:** Data collection and sampling. **Liang Zhao:** Conceiving and designing, reviewing and editing. **Li Zhang:** Original draft preparation. **Zhiqiang Dong:** Visualization, Investigation.

All authors have read and agreed to the submitted version of the manuscript. All authors have read and agreed to the writing of the manuscript.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2024.01.014.

References

[1] Adler DA, Ben-Zeev D, Tseng VW, Kane JM, Brian R, Campbell AT, et al. Predicting early warning signs of psychotic relapse from passive sensing data: an approach using encoder-decoder neural networks. JMIR mHealth uHealth 2020;8:e19962.

[2] Afzal S, Asim M, Javed AR, Beg MO, Baker T. Urldeepdetect: a deep learning approach for detecting malicious urls using semantic vector models. J Netw Syst Manag 2021;29:1–27.

[3] Ahir BK, Engelhard HH, Lakka SS. Tumor development and angiogenesis in adult brain tumor: glioblastoma. Mol Neurobiol 2020;57:2461–78.

[4] Atkinson E, Dickman R. Growth factors and their peptide mimetics for treatment of traumatic brain injury. Bioorg Med Chem 2023:117368.

[5] Bathke J, Lühken G. Ovarflow: a resource optimized gatk 4 based open source variant calling workflow. BMC Bioinforma 2021;22(1):18.

[6] R. Bommasani K. Davis C. Cardie Interpreting pretrained contextualized representations via reductions to static embeddings : Proc 58th Annu Meet Assoc Comput Linguist 2020 4758 4781.

[7] Cavalcanti DD, Kalani MYS, Martirosyan NL, Eales J, Spetzler RF, Preul MC. Cerebral cavernous malformations: from genes to proteins to disease. J Neurosurg 2012;116:122–32.

[8] Chen B, Herten A, Saban D, Rauscher S, Radbruch A, Schmidt B, et al. Hemorrhage from cerebral cavernous malformations: the role of associated developmental venous anomalies. Neurology 2020;95:e89–96.

[9] Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one fastq preprocessor. Bioinformatics 2018;34:i884–90.

[10] Consortium IH, et al. A second generation human haplotype map of over 3.1 million snps. Nature 2007;449:851.

[11] Cui H, Wang C, Maan H, Pang K, Luo F, Wang B. scgpt: Towards building a foundation model for single-cell multi-omics using generative ai. bioRxiv 2023. 2023–04.

[12] Delaneau O, Marchini J. Integrating sequence and array data to create an improved 1000 genomes project haplotype reference panel. Nat Commun 2014;5:3934.

[13] J. Devlin, M.W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding arXiv preprint arXiv 2018.1810.04805.

[14] Gasser E, Sancar G, Downes M, Evans RM. Metabolic messengers: fibroblast growth factor 1. Nat Metab 2022;4:663–71.

[15] Glusman G, Caballero J, Mauldin DE, Hood L, Roach JC. Kaviar: an accessible system for testing snv novelty. Bioinformatics 2011;27:3216–7.

[16] Gu Z, Hübschmann D. simplifyenrichment: a bioconductor package for clustering and visualizing functional enrichment results. Genom Proteom Bioinform 2023;21: 190–202.

[17] Hariri RH, Fredericks EM, Bowers KM. Uncertainty in big data analytics: survey, opportunities, and challenges. J Big Data 2019;6(1):16.

[18] Hashimoto T, Lam T, Boudreau NJ, Bollen AW, Lawton MT, Young WL. Abnormal balance in the angiopoietin-tie2 system in human brain arteriovenous malformations. Circ Res 2001;89:111–3.

[19] Hatva E, Jääskeläinen J, Hirvonen H, Alitalo K, Haltia M. Tie endothelial cell-specific receptor tyrosine kinase is upregulated in the vasculature of arteriovenous malformations. J Neuropathol Exp Neurol 1996;55:1124–33.

[20] Hernandez FG, Carter SJ, Iso-Sipilä J, Goldsmith P, Almousa AA, Gastine S, et al. An automated approach to identify scientific publications reporting pharmacokinetic parameters. Wellcome Open Res 2021;6.

[21] Hirschi KK, Rohovsky SA, D'Amore PA. Pdgf, tgf-β, and heterotypic cell–cell interactions mediate endothelial cell–induced recruitment of 10t1/2 cells and their differentiation to a smooth muscle fate. J Cell Biol 1998;141:805–14.

[22] Ionita-Laza I, McCallum K, Xu B, Buxbaum JD. A spectral approach integrating functional genomic annotations for coding and noncoding variants. Nat Genet 2016;48:214–20.

[23] Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature 2020;581:434–43.

[24] Landrum MJ, Chitipiralla S, Brown GR, Chen C, Gu B, Hart J, et al. Clinvar: improvements to accessing data. Nucleic Acids Res 2020;48:D835–44.

[25] Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. Biobert: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020; 36:1234–40.

[26] Li X, He Y, Wu J, Qiu J, Li J, Wang Q, et al. A novel pathway mutation perturbation score predicts the clinical outcomes of immunotherapy. Brief Bioinform 2022;23: bbac360.

[27] Lipok M, Szlachcic A, Kindela K, Czyrek A, Otlewski J. Identification of a peptide antagonist of the fgf 1–fgfr 1 signaling axis by phage display selection. FEBS Open Bio 2019;9:914–24.

[28] Liu X, Jian X, Boerwinkle E. dbnsfp: a lightweight database of human nonsynonymous snps and their functional predictions. Hum Mutat 2011;32:894–9.

[29] Lopez-Ramirez MA, Lai CC, Soliman SI, Hale P, Pham A, Estrada EJ, et al. Astrocytes propel neurovascular dysfunction during cerebral cavernous malformation lesion formation. J Clin Investig 2021;131.

[30] Mills RE, Pittard WS, Mullaney JM, Farooq U, Creasy TH, Mahurkar AA, et al. Natural genetic variation caused by small insertions and deletions in the human genome. Genome Res 2011;21:830–9.

[31] Mondejar R, Lucas M. Molecular diagnosis in cerebral cavernous malformations. Neurologia 2017;32:540–5.

[32] Moya ML, Cheng MH, Huang JJ, Francis-Sedlak ME, Kao Sw, Opara EC, et al. The effect of fgf-1 loaded alginate microbeads on neovascularization and adipogenesis in a vascular pedicle model of adipose tissue engineering. Biomaterials 2010;31: 2816–26.

[33] Muraina, I., 2022. Ideal dataset splitting ratios in machine learning algorithms: general concerns for data scientists and data analysts, in: 7th International Mardin Artuklu Scientific Research Conference.

[34] Murakami M, Sakurai T. Role of fibroblast growth factor signaling in vascular formation and maintenance: orchestrating signaling networks as an integrated system. Wiley Interdiscip Rev: Syst Biol Med 2012;4:615–29.

[35] Ornitz DM, Itoh N. The fibroblast growth factor signaling pathway. Wiley Interdiscip Rev: Dev Biol 2015;4:215–66.

[36] Orsenigo F, Conze LL, Jauhiainen S, Corada M, Lazzaroni F, Malinverno M, et al. Mapping endothelial-cell diversity in cerebral cavernous malformations at single-cell resolution. Elife 2020;9:e61413.

[37] Ouyang J, Yu H, et al. Natural language description generation method of intelligent image internet of things based on attention mechanism. Secur Commun Netw 2022;2022.

[38] Padarti A, Amritphale A, Eliyas JK, Rigamonti D, Zhang J. Readmissions in patients with cerebral cavernous malformations (ccms): a national readmission database (nrd) study. medRxiv 2021. 2021–09.

[39] Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the american college of medical genetics and genomics and the association for molecular pathology. Genet Med 2015;17:405–23.

[40] Roslin NM, Weili L, Paterson AD, Strug LJ. Quality control analysis of the 1000 genomes project omni2. 5 genotypes. BioRxiv 2016:078600.

[41] Scimone C, Bramanti P, Ruggeri A, Donato L, Alafaci C, Crisafulli C, et al. Ccm3/ serpini1 bidirectional promoter variants in patients with cerebral cavernous malformations: a molecular and functional study. BMC Med Genet 2016;17(1):7.

[42] Scimone C, Bramanti P, Ruggeri A, Katsarou Z, Donato L, Sidoti A, et al. Detection of novel mutation in ccm3 causes familial cerebral cavernous malformations. J Mol Neurosci 2015;57:400–3.

[43] Scimone C, Donato L, Alafaci C, Granata F, Rinaldi C, Longo M, et al. High-throughput sequencing to detect novel likely gene-disrupting variants in pathogenesis of sporadic brain arteriovenous malformations. Front Genet 2020;11: 146.

[44] Scimone C, Donato L, Alibrandi S, Esposito T, Alafaci C, D'Angelo R, et al. Transcriptome analysis provides new molecular signatures in sporadic cerebral cavernous malformation endothelial cells. Biochim Et Biophys Acta (BBA)-Mol Basis Dis 2020;1866:165956.

[45] Sellers F, Palacios-Marqués A, Moliner B, Bernabeu R. Uterine arteriovenous malformation. Case Rep 2013;2013. bcr2012008443.

[46] Songhet P, Adzic D, Reibe S, Rohr KB. fgf1 is required for normal differentiation of erythrocytes in zebrafish primitive hematopoiesis. Dev Dyn: Publ Am Assoc Anat 2007;236:633–43.

[47] Spiegler S, Rath M, Hoffjan S, Dammann P, Sure U, Pagenstecher A, et al. First large genomic inversion in familial cerebral cavernous malformation identified by whole genome sequencing. Neurogenetics 2018;19:55–9.

[48] Srivastava P, Bej S, Yordanova K, Wolkenhauer O. Self-attention-based models for the extraction of molecular interactions from biological texts. Biomolecules 2021; 11:1591.

[49] Storer, K.P., 2006. Cerebral arteriovenous malformations: molecular biology and enhancement of radiosurgical treatment. Ph.D. thesis. UNSW Sydney.

[50] Su VL, Calderwood DA. Signalling through cerebral cavernous malformation protein networks. Open Biol 2020;10:200263.

[51] Targ, S., Almeida, D., Lyman, K., 2016. Resnet in resnet: Generalizing residual architectures. arXiv preprint arXiv:1603.08029.
[52] Uchida S, Shumyatsky GP. Epigenetic regulation of fgf1 transcription by crtc1 and memory enhancement. Brain Res Bull 2018;141:3–12.
[53] Uffelmann E, Huang QQ, Munung NS, DeVries J, Okada Y, Martin AR, et al. Genomewide association studies. Nat Rev Methods Prim 2021;1:59.
[54] de Vos IJ, Vreeburg M, Koek GH, van Steensel MA. Review of familial cerebral cavernous malformations and report of seven additional families. Am J Med Genet Part A 2017;173:338–51.
[55] Wang, S., Scells, H., Koopman, B., Zuccon, G., 2023. Can chatgpt write a good boolean query for systematic review literature search? arXiv preprint arXiv: 2302.03495.
[56] Weng J, Yang Y, Song D, Huo R, Li H, Chen Y, et al. Somatic map3k3 mutation defines a subclass of cerebral cavernous malformation. Am J Hum Genet 2021;108: 942–50.
[57] Wu X, Duan J, Pan Y, Li M. Medical knowledge graph: data sources, construction, reasoning, and applications. Big Data Min Anal 2023;6:201–17.
[58] Xue W, Liu XW, Lee N, Liu QJ, Li WN, Tao H, et al. Features of a chinese family with cerebral cavernous malformation induced by a novelccm1gene mutation. Chin Med J 2013;126:3427–32.
[59] Yang X, Xu W, Leng D, Wen Y, Wu L, Li R, et al. Exploring novel disease-disease associations based on multi-view fusion network. Comput Struct Biotechnol J 2023;21:1807–19.
[60] Yoneyama T, Kasuya H, Onda H, Akagawa H, Jinnai N, Nakajima T, et al. Association of positional and functional candidate genes fgf1, fbn2, and lox on 5q31 with intracranial aneurysm. J Hum Genet 2003;48:309–14.
[61] Zhang B, Qin J. Linc00659 exacerbates endothelial progenitor cell dysfunction in deep vein thrombosis of the lower extremities by activating dnmt3a-mediated fgf1 promoter methylation. Thromb J 2023;21(1):17.
[62] Zhou J, Asteris PG, Armaghani DJ, Pham BT. Prediction of ground vibration induced by blasting operations through the use of the bayesian network and random forest models. Soil Dyn Earthq Eng 2020;139:106390.
[63] Zou Y, Hu J, Huang W, Ye S, Han F, Du J, et al. Non-mitogenic fibroblast growth factor 1 enhanced angiogenesis following ischemic stroke by regulating the sphingosine-1-phosphate 1 pathway. Front Pharmacol 2020;11:59.