Frontiers in Bioscience Landmark

Front. Biosci. (Landmark Ed) **2024**; 29(2): 75
https://doi.org/10.31083/j.fbl2902075

Original Research

Deep-Learning Uncovers certain CCM Isoforms as Transcription Factors

Jacob Croft[1,†], Liyuan Gao[2,†], Victor Sheng[2], Jun Zhang[1,*]

[1]Department of Molecular & Translational Medicine (MTM), Texas Tech University Health Science Center El Paso (TTUHSCEP), El Paso, TX 79905, USA
[2]Department of Computer Sciences, Texas Tech University, Lubbock, TX 79409, USA
*Correspondence: jun.zhang2000@gmail.com (Jun Zhang)
[†]These authors contributed equally.

Academic Editor: Ilaria Baglivo
Submitted: 28 November 2023 Revised: 4 January 2024 Accepted: 22 January 2024 Published: 21 February 2024

Abstract

Background: Cerebral Cavernous Malformations (CCMs) are brain vascular abnormalities associated with an increased risk of hemorrhagic strokes. Familial CCMs result from autosomal dominant inheritance involving three genes: *KRIT1* (*CCM1*), *MGC4607* (*CCM2*), and *PDCD10* (*CCM3*). CCM1 and CCM3 form the CCM Signal Complex (CSC) by binding to CCM2. Both CCM1 and CCM2 exhibit cellular heterogeneity through multiple alternative spliced isoforms, where exons from the same gene combine in diverse ways, leading to varied mRNA transcripts. Additionally, both demonstrate nucleocytoplasmic shuttling between the nucleus and cytoplasm, suggesting their potential role in gene expression regulation as transcription factors (TFs). Due to the accumulated data indicating the cellular localization of CSC proteins in the nucleus and their interaction with progesterone receptors, which serve dual roles as both cellular signaling components and TFs, a question has arisen regarding whether CCMs could also function in both capacities like progesterone receptors. **Methods**: To investigate this potential, we employed our proprietary deep-learning (DL)-based algorithm, specifically utilizing a biased-Support Vector Machine (SVM) model, to explore the plausible cellular function of any of the CSC proteins, particularly focusing on *CCM* gene isoforms with nucleocytoplasmic shuttling, acting as TFs in gene expression regulation. **Results**: Through a comparative DL-based predictive analysis, we have effectively discerned a collective of 11 isoforms across all CCM proteins (CCM1-3). Additionally, we have substantiated the TF functionality of 8 isoforms derived from CCM1 and CCM2 proteins, marking the inaugural identification of CCM isoforms in the role of TFs. **Conclusions**: This groundbreaking discovery directly challenges the prevailing paradigm, which predominantly emphasizes the involvement of CSC solely in endothelial cellular functions amid various potential cellular signal cascades during angiogenesis.

Keywords: cerebral cavernous malformations; CCM2 isoforms; transcription factors; deep-learning; Evolutionary Scale Modeling; Large Language Model; Support Vector Machine; Biased SVM model

1. Introduction

Familial Cerebral Cavernous Malformations (fCCMs) are brain vascular abnormalities with dilated capillaries and increased risk of hemorrhagic strokes. They result from autosomal dominant inheritance with three known genes: *KRIT1* (*CCM1*), *MGC4607* (*CCM2*), and *PDCD10* (*CCM3*). CCM1 and CCM3 proteins bind to CCM2, forming the CCM Signal Complex (CSC) [1]. The complex exhibits nucleocytoplasmic shuttling, initially observed in CCM1. CCM1 possesses nuclear localization signals (NLS) and a nuclear export signal (NES), enabling it to shuttle ICAP1α, a β1 integrin signaling modulator, between the nucleus and cytoplasm [2]. The debate continues on whether CCM1 binds to ICAP1α to direct its localization to the cytoplasmic membrane for inside-out regulation of β1 integrin signaling or vice versa [3]. However, there is a consensus that CCM1 exhibits nucleocytoplasmic shuttling, which modulates the cellular functions of ICAP1α and contributes to its stabilization [4]. CCM1 has a noteworthy impact on the stability of CCM2. Evidence suggests that CCM1 can bind to either ICAP1α or CCM2, facilitating their nuclear-cytoplasmic shuttling and ensuring their stability [5,6]. Initially, lacking any NES and NLS, CCM2 was found to undergo nucleocytoplasmic shuttling with CCM1, similar to ICAP1α, in executing its cellular functions [5,6]. However, recent findings indicate that CCM2 possesses both NLS and NES, suggesting the potential for independent nucleocytoplasmic shuttling without relying on CCM1 [1].

The CSC exhibits notable heterogeneity, with numerous alternatively spliced isoforms documented in the *CCM* genes [1,7]. These isoforms have been shown to undergo nucleocytoplasmic shuttling and play diverse cellular functions [1,8]. While the CSC's function has primarily been studied in cellular signaling and signal transduction pathways, recent reports indicate its involvement in steroid signaling cascades. Experiments revealed that CCM1 protein undergoes nucleocytoplasmic shuttling along with proges-

terone receptors [9]. Given the dual roles of steroid receptors as cellular signaling components and transcription factors, the nuclear localization of CCM1 and CCM2 proteins, their heterogeneity from multiple isoforms, and their interaction with other transcriptional factors, further investigation is needed to understand their potential dual roles as transcription factors.

The nuclear localization of CCM1 and CCM2 proteins, their presence in various isoforms, and their interaction with other transcriptional factors have been well-established over time. Recent reports have expanded our understanding by revealing their involvement in steroid signaling cascades. Notably, new evidence indicates that CCM1 protein undergoes nucleocytoplasmic shuttling in conjunction with progesterone receptors. Despite these advancements, research on the function of CCM proteins has predominantly focused on cellular signaling and signal transduction pathways. Given the dual roles of steroid receptors, serving both as cellular signaling components and transcription factors, a more in-depth exploration is warranted. Further exploration into discerning the nuclear localization of CCM1 and CCM2 proteins, the variety of isoforms they exhibit, and their interactions with other transcriptional factors is crucial for revealing their potential dual roles akin to progesterone receptors. This inquiry, supported by artificial intelligence (AI)-assisted tools developed by our group, is designed to delve into these facets. The results anticipated from this project are expected to markedly improve our understanding of the intricate functions of CCM proteins in cellular processes.

2. Methods

2.1 Experimental Design

This study will employ various deep-learning (DL) based algorisms for comparative analysis for their accuracy and effectiveness of their protein prediction models based on the FASTA, a text-based format for peptide sequences, in which amino acids are represented using single-letter codes. We performed two comparable DL-based algorisms, firstly the commonly used Convolutional Neural Network (CNN) model, then followed by our developed biased-SVM model which involves a two-step process. In biased-SVM model, we firstly use Evolutionary Scale Modeling (ESM), a sophisticated Large Language Model (LLM) for protein sequence representation. Then, we employ a cost-sensitive Support Vector Machine (SVM) classifier to accurately predict transcription factors. Evolutionary Scale Modeling (ESM) is an advanced transformer-based Large Language Model (LLM) designed for proteins, trained on 250 million protein sequences [10]. It predicts protein structure, function, and properties directly from individual sequences using masked language modeling. ESM's flexibility is well-suited for various protein-related tasks. In our methodology, ESM-2 is utilized, a state-of-the-art protein language model that outperforms other single-sequence protein models in structure prediction tasks, enabling atomic resolution predictions of interactions and functions [11]. To handle imbalanced datasets, we use a cost-sensitive Support Vector Machine (SVM) classifier. Traditional SVMs struggle with imbalanced datasets due to their lack of consideration for misclassification costs of minority classes. The cost-sensitive SVM approach assigns misclassification costs inversely proportional to class frequency, ensuring fair treatment of minority classes [12].

This research will employ diverse deep-learning (DL) algorithms to conduct a comparative analysis of the accuracy and effectiveness of protein prediction models based on FASTA, a text-based format representing peptide sequences with single-letter codes for amino acids. Two comparable DL algorithms for transcription factors (TFs) were employed: firstly, the widely-used Convolutional Neural Network (CNN) model, followed by our developed biased-SVM model, which follows a two-step process. In the biased-SVM model, the initial step involves the use of Evolutionary Scale Modeling (ESM), an advanced Large Language Model (LLM) tailored for protein sequence representation. Subsequently, a cost-sensitive Support Vector Machine (SVM) classifier is applied to precisely predict transcription factors. ESM is a sophisticated transformer-based LLM designed for proteins, trained on 250 million protein sequences, providing predictions for protein structure, function, and properties directly from individual sequences using masked language modeling. ESM's adaptability suits various protein-related tasks. In our methodology, ESM-2 is employed, representing a state-of-the-art protein language model that surpasses other single-sequence protein models in structure prediction tasks, enabling atomic resolution predictions of interactions and functions. To address imbalanced datasets, a cost-sensitive Support Vector Machine (SVM) classifier is utilized. Traditional SVMs encounter challenges with imbalanced datasets due to their lack of consideration for misclassification costs of minority classes. The cost-sensitive SVM approach assigns misclassification costs inversely proportional to class frequency, ensuring equitable treatment of minority classes.

2.2 Data Preprocessing

In this research, a robust evaluation methodology is employed, utilizing a 5-fold cross-validation approach on a dataset consisting of 4330 protein sequences. This dataset is comprised of 3315 Non-TF samples and 1015 TF samples. To ensure uniformity and comparability, the sequences are either truncated or padded to a standardized length of 1000 amino acid residues. The performance evaluation of our developed biased-SVM model encompasses various key metrics. Notably, the average macro F1 score, specificity, sensitivity, and balanced accuracy consistently surpass the required significance levels, achieving values of 0.9505, 0.9625, 0.9655, and 0.9640, respectively. These metrics indicate the model's efficacy in accurately predict-

ing transcription factors. Comparatively, the widely-used Convolutional Neural Network (CNN) model, a benchmark in the field, is outperformed by our biased-SVM model. This superiority is evident in the performance metrics (**Supplementary Tables 1,2**), affirming the enhanced predictive capabilities of our developed model across various evaluation criteria.

3. Results

In our previous work, we utilized a machine-learning algorithm to develop a two-step procedure for predicting transcription factors. Firstly, we employed the Evolutionary Scale Modeling (ESM), a protein Large Language Model (LLM), to represent protein sequences [11]. Then, we utilized a cost-sensitive Support Vector Machine (SVM) classifier to efficiently predict transcription factors [11,12]. Our innovative DL-assisted method, referred to as the Biased-SVM model, has been previously reported and demonstrated remarkable accuracy and efficiency [13]. In various previous projects, it has consistently outperformed a Convolutional Neural Network (CNN) model in terms of specificity, sensitivity, and balanced accuracy [11,12]. As mentioned, we used a 5-fold cross-validation on a dataset of 4330 protein sequences, consisting of 3315 Non-Transcription Factor (NTF) samples and 1015 Transcription Factor (TF) samples. To enhance TF prediction, we adopted a lower threshold of 0.1, deviating from the commonly used threshold of 0.5, ensuring no potential TFs were overlooked during the AI-assisted identification process. All sequences were standardized to a length of 1000 amino acid residues through truncation or padding to ensure sequence uniformity, as described in the **Supplementary Materials**. We developed a two-step procedure for TF prediction using machine learning, leveraging ESM, a protein LLM, for sequence representation, and employing a cost-sensitive SVM classifier. The Biased-SVM model outperformed a CNN model in specificity, sensitivity, and balanced accuracy.

Our findings reveal that the CNN model successfully recognized 9 distinct FASTA sequences as transcription factors (TFs) out of the 33 tested sequences, encompassing all *CCM1*, *CCM2*, and *CCM3* genes (Fig. 1A). Subsequently, our optimized Biased-SVM model defined an additional set of 9 isoforms, with 7 of these sequences overlapping with the CNN model (Fig. 1B). These shared sequences—CCM1-Isoform3, CCM2-Isoform116, CCM2-Isoform102, CCM2-Isoform107, CCM2-Isoform402, CCM2-Isoform215, CCM2-Isoform217, CCM2-Isoform609, and CCM2-Isoform213—can be confirmed as TFs in the nucleus.

Furthermore, two isoforms, CCM2-Isoform101 and CCM3-Isoform1, were exclusively predicted by the CNN model and not validated in our Biased-SVM model. This suggests that their potential TF functionality needs further assessment. Additionally, the Biased-SVM model identified two novel isoforms, CCM2-Isoform116 and CCM2-Isoform213, which were not predicted by the CNN model.

To validate the accuracy of our DL-assisted TF identification algorithm, a genetic truncation test was performed on CCM2-Isoform116. Remarkably, by removing the first 70 amino acids and designating it as CCM2-Isoform116 (-N70) in the sequence pool, both CCM2-Isoform116 and CCM2-Isoform116 (-N70) were identified once again as TFs. This result not only confirms the TF functionality of CCM2-Isoform116 but also underscores the robustness of our Biased-SVM model compared to the traditional CNN model (Fig. 1).

4. Discussion

In this study, our Convolutional Neural Network (CNN) model accurately identified 9 distinct FASTA sequences as Transcription Factors (TFs) from the 33 tested Cerebral Cavernous Malformation (CCM) isoforms, covering all known isoforms of *CCM1*, *CCM2*, and *CCM3* genes (Fig. 1A). Our optimized Biased-Support Vector Machine (Biased-SVM) model validated the TF status of 7 isoforms and revealed two previously unidentified isoforms, CCM2-Isoform116 and CCM2-Isoform213. A truncation experiment on CCM2-Isoform116, removing the first 70 amino acids (resulting in CCM2-Isoform116 (-N70)), confirmed both the original and truncated variants as Transcription Factors, providing robust validation for the Biased-SVM model's accuracy (Fig. 1B). Our AI-assisted TF prediction approach offers initial evidence supporting the role of specific CCM isoforms as TFs, challenging existing data and paving the way for nuanced exploration of transcriptional regulatory functions in this domain.

4.1 Utilizing Artificial Intelligence (AI) for Deciphering Biological Complexity

Challenges in big data and multi-omic perspectives on biological complexity, emphasizing the integration of AI for insights into complex biological processes. Biotechnological advancements generate extensive omics data, requiring efficient AI-based algorithms for meaningful insights, with integrated analysis and computational modeling addressing these challenges.

4.1.1 Challenges in Big Data Arising from Biotechnological Advancements in Multi-Omic Perspectives on Biological Complexity

Biological systems function through complex interactions and diverse physiological processes that encompass various types of biomolecules such as DNAs, RNAs, proteins, and gene expression regulatory elements. Each element within this intricate system possesses unique profiles across different omics data types. Achieving a comprehensive understanding of these systems requires an integrated, multi-omic perspective. This emphasizes the importance of integration approaches capable of capturing the complex,

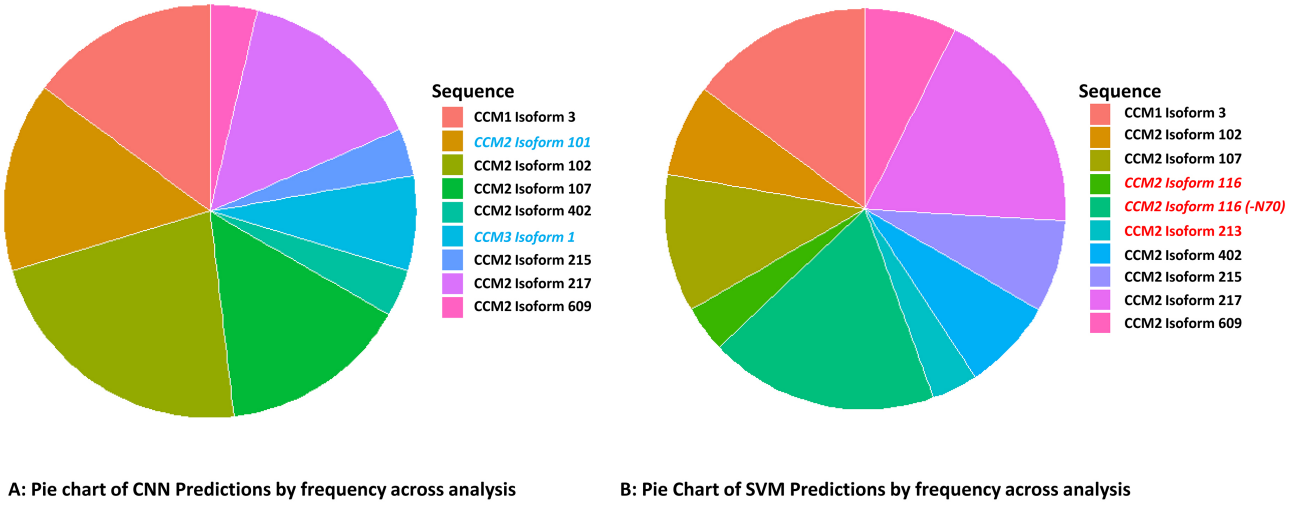

Fig. 1. Multiple machine learning models found concurrent and unique proteins predicted to be transcription factors in multiple repetitions. (A) A pie chart was used to visually represent the predicted transcription factors for each sequence. When considering both sensitivities, the Convolutional Neural Network (CNN) model identified 9 unique FASTA sequences as transcription factors out of the 33 sequences tested. Two of the nine identified isoforms could not be replicated in our biased-Support Vector Machine (SVM) model (highlighted in light blue), raising doubts about their transcription factor (TF) status. (B) A pie chart was utilized to illustrate the biased-SVM model identified 9 distinct sequences as TFs. The partial sequence truncated form of one of the two newly identified isoforms was re-tested and successfully validated (highlighted in red), demonstrating the specificity of our biased-SVM model. The size of each section in the chart corresponds to the number of testing repetitions in which the respective isoform was predicted to be a transcription factor.

often non-linear interactions that define biological systems and addressing the challenges associated with amalgamating heterogeneous data across various omics views [14–18].

4.1.2 Innovative Technologies Unveil Opportunities in Deciphering Biological Complexity

Continual emergence of innovative omics technologies, propelled by biotechnological advancements, enables researchers to access multi-layered information from various biological components. Given the high correlation among these bio-entities, integrative analysis of diverse omics data, encompassing the genome, epigenome, transcriptome, proteome, metabolome, and more, is essential. Multi-omics analysis facilitates the systematic exploration of molecular information at each biological layer, offering a powerful tool for understanding complex biological processes. However, this approach poses challenges in handling the exponentially increasing volume of multi-omics data. Efficient algorithms are crucial to reduce data dimensionality while uncovering the intricacies behind cancer's complex biological processes.

The innovation in biotechnologies, particularly the availability of diverse high-throughput technologies, has resulted in a burgeoning accumulation of omics data, marking the current era as one of 'big data' in biology. Extracting meaningful knowledge from these vast omics datasets remains a formidable challenge in bioinformatics. Therefore, there is an urgent need for improved solutions and innovative methods to efficiently handle and derive meaningful insights from these enormous datasets. Recent advancements in integrated analysis and computational modeling of multi-omics data have started to illuminate this challenging aspect.

4.2 Artificial Intelligence (AI)-Driven Models in Understanding Biological Complexity

AI-based Machine learning and Deep Learning's role in deciphering genomic and proteomic data, showcasing potential in diverse proteomic analyses. The convergence of AI and biology holds promise for unraveling complex biological phenomena.

4.2.1 AI-Driven Models Unlocking Solutions for Complex Biological Inquiries

Recent progress in the integration of AI-driven models and omics has brought together the realms of AI and biology research. This convergence allows for the extraction of patterns from expanding biological databases, curated annotations, or a combination of both. Once these patterns are assimilated, they can be leveraged to yield innovative insights into mechanistic biology and the design of biomolecules.

AI involves the development of intelligent computer programs capable of emulating human intelligence in sensing, reasoning, acting, and adapting. Machine Learning (ML) represents the initial subset of AI-driven algorithms

and applications, with Deep Learning (DL) emerging as the subsequent phase of ML [19,20]. DL utilizes extensive datasets and intricate algorithms to train models, distinguishing itself from ML by employing neural networks with a level of complexity that mirrors the human brain—referred to as Deep Neural Networks (DNNs). DNNs enhance AI capabilities by adeptly processing complex biological data and predicting diverse molecular activities measured through high-throughput functional Omics assays. These networks have the capacity to unveil patterns, such as sequence motifs of large biomolecules. In practice, DL has demonstrated its proficiency in analyzing complementary multi-modal data streams, while AI has showcased its ability to provide decisive interpretations of voluminous and intricate datasets. Consequently, AI-based analysis emerges as the most effective tool for unraveling biological complexity [21–24].

4.2.2 Application of DL-Based Algorithms for Omics

DL-based Models have played a substantial role in understanding the regulatory code that governs gene expression in genomic data. This encompasses the investigation of diverse elements involved in gene expression regulation [25], such as transcription factor binding sites [26–28], transcription start sites (TSSs) [28], missing data extrapolations [29], and other related genomic components [25,30]. This understanding has been particularly enhanced by the existence of extensive regulatory sequence datasets that are essential for the "AI-learning" process [14–18].

In contrast to genomics, the interpretation of proteomics, especially the prediction of protein tertiary structure and function from sequence data, has notably lagged behind. This discrepancy is primarily attributed to the delayed progress of proteomic technologies and the inherent uniqueness of the biophysical and biochemical properties among various proteins. Historically, the lack of proteomics data has hindered progress, but in recent years, this trend has undergone a significant reversal due to the widespread adoption of newly developed high-throughput proteomics technologies, resulting in a substantial increase in deposited "protein data sets" [22,31–33].

These breakthrough technologies, contributing to the exponential growth of protein sequence data, now provide a robust foundation for the application of AI-assisted computations, particularly through DL-based models. Processes such as extracting physical contacts from the evolutionary protein record, distilling sequence-structure patterns from known protein structures, incorporating templates from homologs in the Protein Data Bank (PDB), and refining coarsely predicted structures into finely resolved ones have all been redefined using neural networks. Collectively, this transformation has yielded algorithms capable of predicting single protein domains with a median accuracy.

4.2.3 Overcome the Limits of DL-Based Algorithms for Complex Biological Inquiries

DL-based models have demonstrated exceptional performance in interpreting diverse and intricate inquiries. They have been effectively employed in various applications, such as misinformation detection in Cyber-Physical Social Services [34]; traffic safety management and vehicle navigation in intelligent transportation systems [35]; ontology matching to address the semantic heterogeneity problem [36]; online measures for protecting patient health information, ensuring compliance with HIPAA in healthcare systems [37], and the development of quick and reliable diagnosis tools for early-stage detection and treatment of liver cancer [38].

When utilizing either machine learning or more advanced deep learning models for predictions in AI-assisted analysis, they are initially trained on a set of known data. Subsequently, based on this training, the "trained" model attempts to predict outcomes for new data sets. However, a common error that may occur during this process is overfitting, where the model can provide accurate predictions for the training data but not for new data. An overfit model may produce inaccurate predictions and may not perform well with various types of new data. To address this critical and potential limitation, we performed a comparative predictive analysis based on deep learning. Specifically, we simultaneously employed two DL-based TF prediction models: the Convolutional Neural Network (CNN) model and our newly developed Large Language Model (LLM), the biased-SVM model, as outlined below. The congruent outcomes from both models (Fig. 1) not only validate our TF predictions but also reduce the risk of overfitting. Moreover, to continue minimizing the risk of overfitting linked to DL models, our plan is to consistently expand our pool of known TFs. Following this, we will engage in both *in-vitro* and *in-vivo* experiments to reevaluate the cellular functions of these TFs within CCM isoforms.

As previously noted, DL-based models exhibit potential in predicting functional genomics and can generate protein sequences with foreseeable functions across extensive protein families. Furthermore, the utilization of DL-based models for predicting 3D structures of large proteins is becoming a noteworthy research area. Beyond their contributions to protein structure and function analysis, DL-based models possess significant potential for diverse facets of proteomic analysis, including enzymatic activities [39], post-translational modification (PTM) status [30], protein classification [15,30], and even design and engineering for novel protein species [20,33].

4.3 A Novel Large Language Model (LLM) Algorithm for Transcription Factor Prediction

To addressing DL model limitations. LLMs, derived from well-trained protein DL models, offer enhanced predictions in protein domains, capturing evolutionary rela-

tionships and facilitating knowledge transfer [40]. The study applies the biased-SVM model, an LLM, to identify CCM isoforms as transcription factors, challenging established paradigms.

4.3.1 An Innovative and DL-Derived Large Language Model (LLM) Algorithm for Transcription Factor Prediction

Recently, there have been growing concerns about the constraints and shortcomings of DL-based models, especially concerning their effectiveness and interpretability. Although recent DL models have demonstrated enhanced accuracy in predicting variant effects, they encounter challenges when analyzing all coding variants in proteins, often depending on close homologs or DL-based algorithms. A notable limitation of DL models is evident in the importance scores within DNN attribution maps, which can sometimes be spurious, introducing uncertainty into the reliability of model selection, even for well-performing DNNs.

To tackle these challenges, there is a need for modified approaches that measure the consistency of crucial features across a population of attribution maps for proteomic tasks. In response, a novel model is introduced, leveraging well-trained protein DL-derived language models, defined as Large Language Model (LLM) [31,40–43]. LLMs, initiated from ensembles of evolutionarily related protein sequences, capture representations of protein families, facilitating the generation of functional protein sequences. When applied to protein sequence data, these models advance predictions related to structure, function, and mutational effects. Given their exceptional performance, LLMs have been seamlessly integrated into common benchtop proteomic domains, such as AlphaFold [44]. These models generate functional protein sequences from ensembles of evolutionarily related protein sequences, capturing representations of protein families and acquiring knowledge about constraints associated with protein structure and function. Pre-trained natural language processing models, fine-tuned for in-domain tasks, transfer learned knowledge from a vast natural language corpus to protein domains. The use of protein language models as foundational structures, pre-trained on extensive natural language corpora, streamlines the handling of protein sequences, enabling them to encapsulate information about protein quaternary states.

4.3.2 Our Innovative Biased-SVM Model Excels over Traditional DL-Based Algorithm in Transcription Factor Prediction

Building on this direction, we utilized our newly developed a Large Language Model (LLM), the biased-SVM model [13], to explore the cellular functions of Cerebral Cavernous Malformation (CCM) proteins. Our focus was specifically on *CCM* gene isoforms with nucleocytoplasmic shuttling, serving as transcription factors (TFs) in the regulation of gene expression. The outcomes validate the enhanced performance of the biased-SVM model compared to the conventional DL-based TF prediction model [13,45–48]. In a comparative DL-based predictive analysis, we identified 11 isoforms across all CCM proteins (CCM1-3) and validated the TF functionality of 8 isoforms derived from CCM1 and CCM2 proteins. This groundbreaking discovery challenges the prevailing paradigm, which primarily emphasizes the involvement of CCM proteins solely in endothelial cellular functions amid various potential cellular signal cascades during angiogenesis. While the phenomena of inside-out and outside-in signaling in integrin signaling of endothelial cells during angiogenesis have long been recognized, the underlying mechanism remains largely unexplored [4,49–51]. Additionally, despite having previously linked CSC signaling and integrin signaling [4], the current discovery that many CCM isoforms act as TFs may represent the missing link required for understanding endothelial cell and angiogenesis dynamics. Nevertheless, addressing this substantial knowledge gap necessitates extensive experiments. To further our understanding of the CCM signaling complex in endothelial cell biology and angiogenesis, it is imperative to validate the identified candidate TFs within CCM isoforms. This validation, particularly in CCM gene knockout cell lines *in vitro* and *CCM* gene knockout mice *in vivo*, is crucial for advancing our comprehension of the intricate processes involved. We are confident that the continuation of this project, along with the implementation of the proposed experiments, will enable a deeper understanding of endothelial cell signaling, potentially shaping the trajectory of angiogenesis. The insights gleaned from this research hold the potential to profoundly impact future strategies for the treatment, prevention, or even the cure of cerebral vascular diseases.

5. Conclusions

This study employed a comparative approach, simultaneously utilizing both the widely adopted deep-learning model, Convolutional Neural Network (CNN), and our optimized Biased-Support Vector Machine (Biased-SVM) models to detect Transcription Factors (TFs) among Cerebral Cavernous Malformation (CCM) isoforms. The CNN accurately pinpointed 9 TFs, covering all *CCM1-3* genes, while the Biased-SVM confirmed 7 TFs from the pool of these 9 CNN-identified TFs and additionally revealed two new isoforms as TFs. Subsequently, through our analysis employing a truncation experimental design, we validated their TF status. Our findings challenge existing data and their associated models, proposing previously undiscovered functions for specific CCM isoforms.

Availability of Data and Materials

Data can be provided by the corresponding author upon request after a 6-month embargo period.

Author Contributions

JC: Methodology, Writing – Original draft preparation, Writing- reviewing and Editing, Experiments: Data collection, data curation and management; LG: Methodology, Writing – Original draft preparations; VS: Conceptualization, Methodology Writing – Original draft preparation, Reviewing and Editing; JZ: Conceptualization, Methodology, Writing – Original draft preparation, Reviewing, Editing, and Finalized manuscript. All authors read and approved the final manuscript. All authors have participated sufficiently in the work and agreed to be accountable for all aspects of the work.

Ethics Approval and Consent to Participate

Not applicable.

Acknowledgment

We wish to thank Muaz Bhalli, Alexander Le, Ofek Belkin, Mellisa Renteria, David Jang, Justin Aickareth, Victoria Reid, Majd Hawwar, Revathi Gnanasekaran, Nickolas Sanchez, Charlie Harvey, and Drexell Vincent at Texas Tech University Health Science Center El Paso (TTUHSCEP) for their technical help during the experiments.

Funding

This research received no external funding.

Conflict of Interest

The authors declare no conflict of interest.

Supplementary Material

Supplementary material associated with this article can be found, in the online version, at https://doi.org/10.31083/j.fbl2902075.

References

[1] Jiang X, Padarti A, Qu Y, Sheng S, Abou-Fadel J, Badr A, *et al*. Alternatively spliced isoforms reveal a novel type of PTB domain in CCM2 protein. Scientific Reports. 2019; 9: 15808.

[2] Abou-Fadel J, Grajeda B, Jiang X, Cailing-De La O AMD, Flores E, Padarti A, *et al*. CmP signaling network unveils novel biomarkers for triple negative breast cancer in African American women. Cancer Biomarkers: Section a of Disease Markers. 2022; 34: 607–636.

[3] Zhang J, Basu S, Rigamonti D, Dietz HC, Clatterbuck RE. Krit1 modulates beta 1-integrin-mediated endothelial cell proliferation. Neurosurgery. 2008; 63: 571–578; discussion 578.

[4] Zhang J, Clatterbuck RE, Rigamonti D, Chang DD, Dietz HC. Interaction between krit1 and icap1alpha infers perturbation of integrin beta1-mediated angiogenesis in the pathogenesis of cerebral cavernous malformation. Human Molecular Genetics. 2001; 10: 2953–2960.

[5] Zhang J, Rigamonti D, Dietz HC, Clatterbuck RE. Interaction between krit1 and malcavernin: implications for the pathogenesis of cerebral cavernous malformations. Neurosurgery. 2007; 60: 353–359; discussion 359.

[6] Faurobert E, Rome C, Lisowska J, Manet-Dupé S, Boulday G, Malbouyres M, *et al*. CCM1-ICAP-1 complex controls β1 integrin-dependent endothelial contractility and fibronectin remodeling. The Journal of Cell Biology. 2013; 202: 545–561.

[7] Retta SF, Avolio M, Francalanci F, Procida S, Balzac F, Degani S, *et al*. Identification of Krit1B: a novel alternative splicing isoform of cerebral cavernous malformation gene-1. Gene. 2004; 325: 63–78.

[8] Francalanci F, Avolio M, De Luca E, Longo D, Menchise V, Guazzi P, *et al*. Structural and functional differences between KRIT1A and KRIT1B isoforms: a framework for understanding CCM pathogenesis. Experimental Cell Research. 2009; 315: 285–303.

[9] Aickareth J, Hawwar M, Sanchez N, Gnanasekaran R, Zhang J. Membrane Progesterone Receptors (mPRs/PAQRs) Are Going beyond Its Initial Definitions. Membranes. 2023; 13: 260.

[10] Rives A, Meier J, Sercu T, Goyal S, Lin Z, Liu J, *et al*. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proceedings of the National Academy of Sciences of the United States of America. 2021; 118: e2016239118.

[11] Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, *et al*. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science (New York, N.Y.). 2023; 379: 1123–1130.

[12] Cao P, Zhao D, Zaïane OR. An Optimized Cost-Sensitive SVM for Imbalanced Data Learning. Springer Berlin Heidelberg EBooks. 2013. Available at: https://link.springer.com/chapter/10.1007/978-3-642-37456-2_24 (Accessed: 1 October 2023).

[13] Gao L, Shu K, Zhang J, Sheng VS. Explainable Transcription Factor Prediction with Protein Language Models. 2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). 2023: 853–856.

[14] He X, Liu X, Zuo F, Shi H, Jing J. Artificial intelligence-based multi-omics analysis fuels cancer precision medicine. Seminars in Cancer Biology. 2023; 88: 187–200.

[15] Kumar K, Bhowmik D, Mandloi S, Gautam A, Lahiri A, Biswas N, *et al*. Integrating Multi-Omics Data to Construct Reliable Interconnected Models of Signaling, Gene Regulatory, and Metabolic Pathways. Methods in Molecular Biology (Clifton, N.J.). 2023; 2634: 139–151.

[16] Biswas N, Chakrabarti S. Artificial Intelligence (AI)-Based Systems Biology Approaches in Multi-Omics Data Analysis of Cancer. Frontiers in Oncology. 2020; 10: 588221.

[17] Picard M, Scott-Boyer MP, Bodein A, Périn O, Droit A. Integration strategies of multi-omics data for machine learning analysis. Computational and Structural Biotechnology Journal. 2021; 19: 3735–3746.

[18] Li R, Li L, Xu Y, Yang J. Machine learning meets omics: applications and perspectives. Briefings in Bioinformatics. 2022; 23: bbab460.

[19] Wu IW, Tsai TH, Lo CJ, Chou YJ, Yeh CH, Chan YH, *et al*. Discovering a trans-omics biomarker signature that predisposes high risk diabetic patients to diabetic kidney disease. NPJ Digital Medicine. 2022; 5: 166.

[20] Mardikoraem M, Wang Z, Pascual N, Woldring D. Generative models for protein sequence modeling: recent advances and future directions. Briefings in Bioinformatics. 2023; 24: bbad358.

[21] Wu F, Wu L, Radev D, Xu J, Li SZ. Integration of pre-trained protein language models into geometric deep learning networks. Communications Biology. 2023; 6: 876.

[22] Buehler MJ. Multiscale Modeling at the Interface of Molecular Mechanics and Natural Language through Attention Neural Networks. Accounts of Chemical Research. 2022; 55: 3387–3403.

[23] Lee NK, Tang Z, Toneyan S, Koo PK. EvoAug: improving generalization and interpretability of genomic deep neural networks with evolution-inspired data augmentations. Genome Biology. 2023; 24: 105.

[24] Majdandzic A, Rajesh C, Tang A, Toneyan S, Labelson E, Tripathy R, *et al*. Selecting deep neural networks that yield consistent attribution-based interpretations for genomics. Proceedings of Machine Learning Research. 2022; 200: 131–149.
[25] Ding K, Dixit G, Parker BJ, Wen J. CRMnet: A deep learning model for predicting gene expression from large regulatory sequence datasets. Frontiers in Big Data. 2023; 6: 1113402.
[26] Shen Z, Bao W, Huang DS. Recurrent Neural Network for Predicting Transcription Factor Binding Sites. Scientific Reports. 2018; 8: 15270.
[27] Kim GB, Gao Y, Palsson BO, Lee SY. DeepTFactor: A deep learning-based tool for the prediction of transcription factors. Proceedings of the National Academy of Sciences of the United States of America. 2021; 118: e2021171118.
[28] Koo PK, Ploenzke M. Deep learning for inferring transcription factor binding sites. Current Opinion in Systems Biology. 2020; 19: 16–23.
[29] Flores JE, Claborne DM, Weller ZD, Webb-Robertson BJM, Waters KM, Bramer LM. Missing data in multi-omics integration: Recent advances through artificial intelligence. Frontiers in Artificial Intelligence. 2023; 6: 1098308.
[30] Pokharel S, Pratyush P, Ismail HD, Ma J, Kc DB. Integrating Embeddings from Multiple Protein Language Models to Improve Protein *O*-GlcNAc Site Prediction. International Journal of Molecular Sciences. 2023; 24: 16000.
[31] AlQuraishi M. Machine learning in protein structure prediction. Current Opinion in Chemical Biology. 2021; 65: 1–8.
[32] Avraham O, Tsaban T, Ben-Aharon Z, Tsaban L, Schueler-Furman O. Protein language models can capture protein quaternary state. BMC Bioinformatics. 2023; 24: 433.
[33] Ferruz N, Heinzinger M, Akdel M, Goncearenco A, Naef L, Dallago C. From sequence to function through structure: Deep learning for protein design. Computational and Structural Biotechnology Journal. 2022; 21: 238–250.
[34] Zhang Q, Guo Z, Zhu Y, Vijayakumar P, Castiglione A, Gupta BB. A Deep Learning-based Fast Fake News Detection Model for Cyber-Physical Social Services. Pattern Recognition Letters. 2023; 168: 31–38.
[35] Liu RW, Guo Y, Lu Y, Chui KT, Gupta BB. Deep Network-Enabled Haze Visibility Enhancement for Visual IoT-Driven Intelligent Transportation Systems. IEEE Transactions on Industrial Informatics. 2023; 19: 1581–1591.
[36] Khoudja MA, Fareh M, Bouarfa H. Deep Embedding Learning with Auto-Encoder for Large-Scale Ontology Matching. International Journal on Semantic Web and Information Systems (IJSWIS). 2022; 18: 18.
[37] Nguyen GN, Viet NHL, Elhoseny M, Shankar K, Gupta BB, El-Latif AAA. Secure blockchain enabled Cyber–physical systems in healthcare using deep belief network with ResNet model. Journal of Parallel and Distributed Computing. 2021; 153: 150–160.
[38] Anil BC, Dayananda P, Nethravathi B, Raisinghani MS. Efficient Local Cloud-Based Solution for Liver Cancer Detection Using Deep Learning. International Journal of Cloud Applications and Computing (IJCAC). 2022; 12: 13.
[39] Xie WJ, Warshel A. Harnessing Generative AI to Decode Enzyme Catalysis and Evolution for Enhanced Engineering. BioRxiv: the Preprint Server for Biology. 2023. (preprint)
[40] Sgarbossa D, Lupo U, Bitbol AF. Generative power of a protein language model trained on multiple sequence alignments. eLife. 2023; 12: e79854.
[41] Madani A, Krause B, Greene ER, Subramanian S, Mohr BP, Holton JM, *et al*. Large language models generate functional protein sequences across diverse families. Nature Biotechnology. 2023; 41: 1099–1106.
[42] Hu M, Alkhairy S, Lee I, Pillich RT, Bachelder R, Ideker T, *et al*. Evaluation of large language models for discovery of gene set function. ArXiv. 2023. (preprint)
[43] Yue T, Wang Y, Zhang L, Gu C, Xue H, Wang W, *et al*. Deep Learning for Genomics: From Early Neural Nets to Modern Large Language Models. International Journal of Molecular Sciences. 2023; 24: 15858.
[44] Chen T, Pertsemlidis S, Watson R, Kavirayuni VS, Hsu A, Vure P, *et al*. PepMLM: Target Sequence-Conditioned Generation of Peptide Binders via Masked Language Modeling. ArXiv. 2023. (preprint)
[45] Croft J, Grajeda B, Abou-Fadel J, Ellis C, Estevao IL, Almeida IC, *et al*. Blood prognostic biomarker signatures for hemorrhagic cerebral cavernous malformations (CCMs). BioRxiv. 2023. (preprint)
[46] Croft J, Grajeda B, Aguirre LA, Gao L, Abou-Fadel J, Sheng V, *et al*. Whole-genome Omics delineates the function of CCM1 within the CmPn networks. BioRxiv. 2023. (preprint)
[47] Croft J, Quintanar O, Zhang J. Updated Biomarkers for TNBC in African vs. Caucasian American Women. BioRxiv. 2023. (preprint)
[48] Croft J, Gao LY, Quintanar O, Sheng V, Zhang J. Identification of Cholangiocarcinoma (CCA) Subtype-Specific Biomarkers. BioRxiv. 2023. (preprint)
[49] Zhang J, Clatterbuck RE, Rigamonti D, Chang DD, Dietz HC. Novel insights regarding the pathogenesis of cerebral cavernous malformation (CCM). American Journal of Human Genetics. 2001; 69: 178.
[50] Liu H, Rigamonti D, Badr A, Zhang J. Ccm1 assures microvascular integrity during angiogenesis. Translational Stroke Research. 2010; 1: 146–153.
[51] Liu H, Rigamonti D, Badr A, Zhang J. Ccm1 regulates microvascular morphogenesis during angiogenesis. Journal of Vascular Research. 2011; 48: 130–140.