

Characterizing and predicting ccRCC-causing missense mutations in Von Hippel-Lindau disease

Adam Serghini¹, Stephanie Portelli¹, Guillaume Troadec², Catherine Song², Qisheng Pan^{1,3}, Douglas E.V. Pires^{2,3},

David B. Ascher^{1,3,*}

¹School of Chemistry and Molecular Biosciences, Chemistry Building 68, Cooper Road, The University of Queensland, St Lucia, QLD 4072, Queensland, Australia

²School of Computing and Information Systems, University of Melbourne, Melbourne, VIC 3010, Australia

³Computational Biology and Clinical Informatics, Baker Heart and Diabetes Institute, 75 Commercial Road, Melbourne, VIC 3004, Australia

*Corresponding author. D.B.A. Chemistry Building 68, Cooper Road, The University of Queensland, St Lucia, QLD 4067, Australia. E-mail: d.ascher@uq.edu.au.

Abstract

Background: Mutations within the Von Hippel-Lindau (VHL) tumor suppressor gene are known to cause VHL disease, which is characterized by the formation of cysts and tumors in multiple organs of the body, particularly clear cell renal cell carcinoma (ccRCC). A major challenge in clinical practice is determining tumor risk from a given mutation in the VHL gene. Previous efforts have been hindered by limited available clinical data and technological constraints. **Methods:** To overcome this, we initially manually curated the largest set of clinically validated VHL mutations to date, enabling a robust assessment of existing predictive tools on an independent test set. Additionally, we comprehensively characterized the effects of mutations within VHL using *in silico* biophysical tools describing changes in protein stability, dynamics and affinity to binding partners to provide insights into the structure-phenotype relationship. These descriptive properties were used as molecular features for the construction of a machine learning model, designed to predict the risk of ccRCC development as a result of a VHL missense mutation. **Results:** Analysis of our model showed an accuracy of 0.81 in the identification of ccRCC-causing missense mutations, and a Matthew's Correlation Coefficient of 0.44 on a non-redundant blind test, a significant improvement in comparison to the previous available approaches. **Conclusion:** This work highlights the power of using protein 3D structure to fully explore the range of molecular and functional consequences of genomic variants. We believe this optimized model will better enable its clinical implementation and assist guiding patient risk stratification and management.

Keywords: Von Hippel-Lindau; VHL; clear cell renal cell carcinoma; ccRCC; machine learning

Introduction

Von Hippel-Lindau (VHL) disease has the inheritance pattern of an autosomal dominant disease, with an estimated prevalence of up to 1 in 36 000 individuals among the general population [1]. This condition can lead to the development of several pathological conditions including cysts and tumors within multiple organs of the body, such as the liver, kidneys, retina, brain and spinal cord [2]. Common oncological manifestations of VHL disease include Clear Cell Renal Cell Carcinoma (ccRCC), pheochromocytoma (PPC), and hemangioblastomas within the central nervous system and retina, though one of the most frequently observed tumor pathologies of VHL disease is ccRCC [3, 4]. VHL disease is brought about through the acquisition of inactivating mutations in both alleles of the VHL gene within a single cell, with approximately 52% of such mutations classified as missense mutations [5]. VHL is classified as a tumor suppressor gene, therefore mutations that inactivate this gene predispose VHL patients to neoplasm generation and subsequent tumor formation [6].

Multiple isoforms of the VHL protein are generated through alternative splicing mechanisms, with the isoform of concern regarding VHL disease being the canonical pVHL isoform 1 protein product (pVHL30). This is due to its role in regulating

HIF1 α (Hypoxia-inducible factor 1-alpha), as well as other proteins involved in tumor suppression [7]. The pVHL30 consists of unstructured regions and two structured domains: the alpha and beta domains. The alpha domain is responsible for forming a complex with elongin B and elongin C, while the beta domain interacts with hydroxylated prolines of HIF-1 α (Supplementary Fig. 1). The physiological significance of the pVHL30 structure varies along different regions of the protein, with the N-terminal domain being shown to be very disordered where augmentations are less likely to impact the molecular functions of the protein, while the alpha and beta domains exhibit a large degree of orderliness and hence are much more likely to play a critical role in the tumor suppressor functions of the protein [8–10].

The VHL protein reduces tumor formation via modification of multiple chemical signaling pathways, the most well studied of which being its role in HIF-1 α regulation. A complex is formed between pVHL, Elongin B and Elongin C, in the cytoplasm of the cell. Via the recruitment of the E3 ubiquitin ligase complex, VHL facilitates the ubiquitination and subsequent proteasomal degradation of HIF-1 α under normal physiological conditions [11, 12]. However, in hypoxic conditions or in the event that pVHL becomes compromised (as observed in VHL disease), the binding

Received: August 20, 2023. Revised: October 19, 2023. Accepted: October 20, 2023

© The Author(s) 2023. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

of pVHL to HIF-1 α does not occur. This leads to the accumulation of HIF-1 α within the cell's cytoplasm and its translocation to the nucleus, where it acts as a transcription factor targeting and upregulating hypoxia-responsive genes. This includes genes involved in anaerobic energy metabolism (e.g. *GLUT1*), regulating oxygen transportation (e.g. *EPO*), and initiating angiogenesis (e.g. *VEGF*), resulting in a higher probability of the formation of neoplasms and subsequent tumor formation [13–16]. pVHL also acts to support the activation of the p53 tumor suppressor protein [17, 18]. However, the interactions between VHL and p53 are not well defined.

Innovations in the field of computational biology have provided highly accurate methodologies to assess the impact of mutations on the risk of disease development [19, 20]. We have previously interrogated this concept using structure-based statistical and machine learning approaches [21, 22] in genetic diseases [23–30] and cancers [31–36], as well as cancer and anti-microbial drug resistance [37–45]. Within the VHL space, we previously developed a machine-learning model, Symphony, capable of predicting the risk of ccRCC development upon missense mutation [46]. While Symphony exhibited good accuracy and specificity when it was first published, it primarily considered effects of mutations on protein thermal stability. The increased amount of publicly available data regarding VHL mutation pathogenicity and advancement of molecular feature generation tools since its inception presents the opportunity to evaluate these previous models, and integrate a more comprehensive picture of molecular and functional consequences of mutations to improve predictive performance. In this work, we built upon findings from Symphony using novel techniques and current clinical data to develop an up-to-date ccRCC development predictor for missense mutations in the VHL gene. Our new tool offers invaluable applications clinically, particularly in prioritizing patient monitoring protocols according to cancer risk.

Results

Data curation

We obtained ccRCC and non-ccRCC VHL mutations from the Symphony database, which consisted of a total of 121 missense mutations, and were used for the training of our model. Additionally, our literature search from 96 unique articles identified a further 92 mutations, which were not considered during the development of Symphony, and were used as a non-redundant test set. Overall, our data consisted of 142 ccRCC-causing mutations and 71 non-ccRCC causing mutations, (Supplementary Table 1).

In observing the distributions of both classes of mutations in a linear (lollipop plot, Fig. 1A) and 3-dimensional representation (Fig. 1B), we observed that mutations are mainly situated within the structured regions of the VHL protein, with multiple overlaps in phenotypes within the same loci. Additionally, there does not seem to be any clear discernible pattern of differentiation between the two phenotypes and mutation location from these illustrations alone.

In order to further investigate the relationship between the mutation position and ccRCC development, we used MTR3D to estimate missense intolerance of pVHL30. MTR3D is a computational tool that estimates the deleteriousness of a missense variant, by utilizing datasets of observed variation within the population to calculate the Missense Tolerance Ratio (MTR). The MTR value is used to identify regions of a protein that are under a selection pressure and that are intolerant to genetic changes [46–48]. Average MTR scores of alpha and beta domains are 0.52 and

Table 1. Comparison of performance of prediction tools for ccRCC development from VHL missense mutations.

Performance metric	Symphony	SIFT	Polyphen-2
MCC	0.075	0.032	0.036

Each predictive tool was evaluated using the MCC metric, using the same non-redundant dataset.

0.65, respectively, suggesting the alpha domain is more tolerant to mutations than the beta domain (Fig. 2). Interestingly, however, there are two small regions of the beta domain where no ccRCC-causing mutations have been recorded (residues 137–148 and 193–203), which map onto a linker region and alpha helix of the domain. This absence of ccRCC-causing mutations is reflected in a lower average MTR score for these regions of 0.44 compared to that of the rest of the beta domain, which has an average score of 0.72.

Evaluation of current pathology prediction tools

In order to allow for a direct comparison of our final model with other known tools, a literature-curated, non-redundant dataset was used as a benchmarking dataset. Prediction tools tested included the VHL-specific tool Symphony, and more general disease predictors, Polyphen-2 and SIFT (see Table 1). PolyPhen-2 (Polymorphism Phenotyping v2) predicts the functional and stability changes of a protein from single nucleotide polymorphisms [49] and SIFT (Sorting Intolerant From Tolerant) is a sequence homology-based tool that predicts a change in the phenotypic effect from a single amino acid change [50]. This analysis demonstrated the poor performance of the aforementioned tools in distinguishing between ccRCC and non-ccRCC causing mutations on this newly curated data set, demonstrating the need for a novel and optimized model to assist on VHL patient risk stratification.

Statistical analysis

In order to further our understanding of VHL-mediated ccRCC, statistically significant (defined as P -value < 0.05) discrepancies of molecular and biophysical characteristics between ccRCC and non-ccRCC VHL missense mutations were identified through a Welch sample t -test.

Interestingly, of all the molecular features tested, the feature that displayed the highest statistical difference between the two phenotypes was generated from a stochastic gradient boosting model called Envision (P -value $< 9.9E-9$) (Fig. 3A). Envision predicts the extent to which a mutation is damaging for a protein compared to the wild-type, providing a quantitative predictor of molecular effect [51].

Furthermore, the likelihood of amino acid change is highly dependent on the disparities of physicochemical properties of the residues, such as residue pH and polarity, thus it was suspected that these features may differ between ccRCC and non-ccRCC mutations. Numerous features describing the evolutionary conservation and amino acid substitution probabilities were generated, and calculated using PAM and BLOSUM matrices, which describe the rate at which amino acids are substituted for other amino acids. Analysis of the sequence-based features revealed noticeable disparities between the two phenotypes, with higher scores observed in the non-tumor causing mutations compared to that of the tumor causing mutations. For example Blosum62 (P -value = $2.36E-3$) as seen in Fig. 3B, indicates that pVHL30 non-ccRCC mutations occur more frequently than ccRCC-causing mutations.

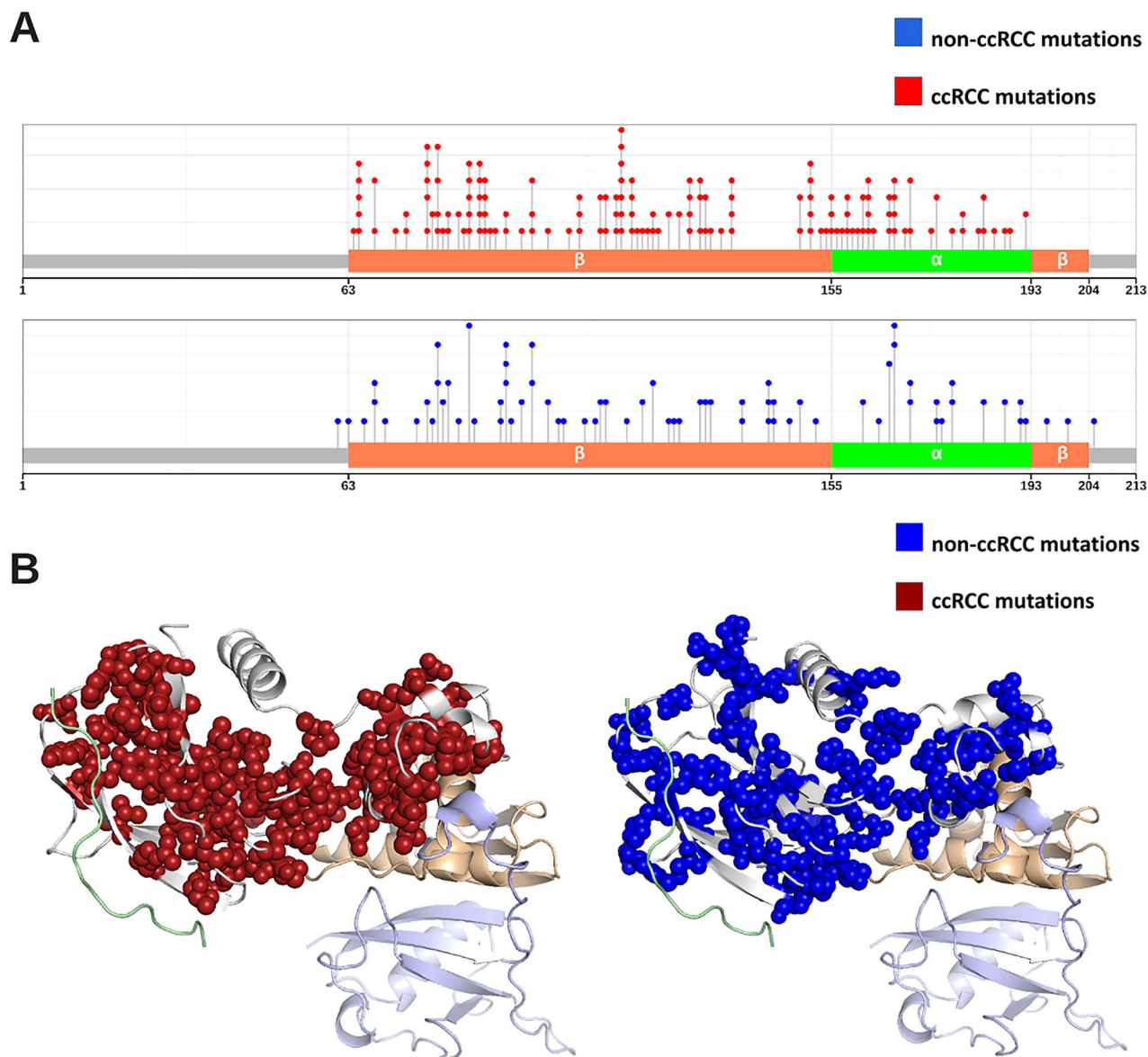


Figure 1. pVHL30 mutation distribution by phenotype. Illustration of the ccRCC and non-ccRCC mutations in the primary sequence of pVHL30 (A), and its 3-dimensional structure (B). Mutations of both phenotypes are concentrated in the structured regions of the alpha and beta domains, and are mostly absent from the unstructured regions of pVHL30.

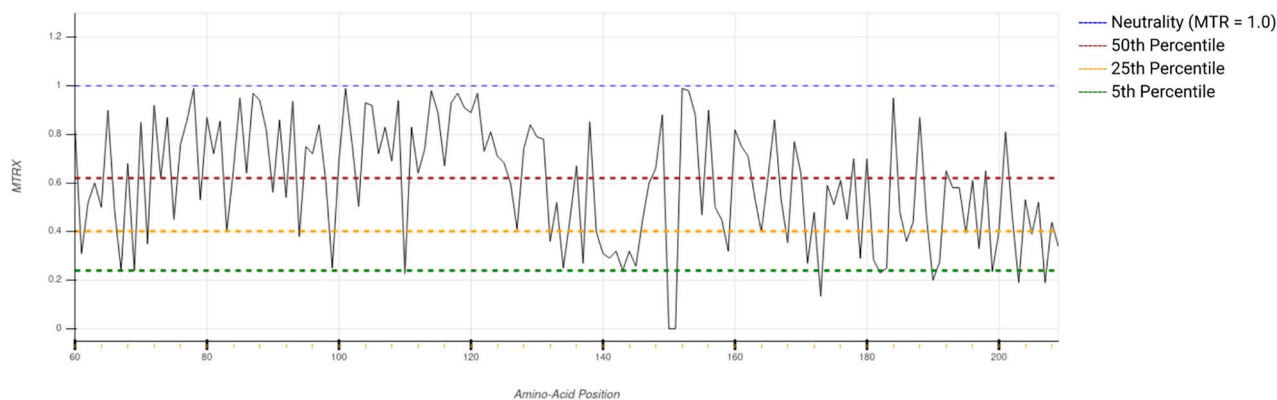


Figure 2. MTR scores of pVHL30 amino-acid position calculated by MTR3D. Horizontal lines show gene-specific MTR percentiles 5th, 25th, 50th, and neutrality.

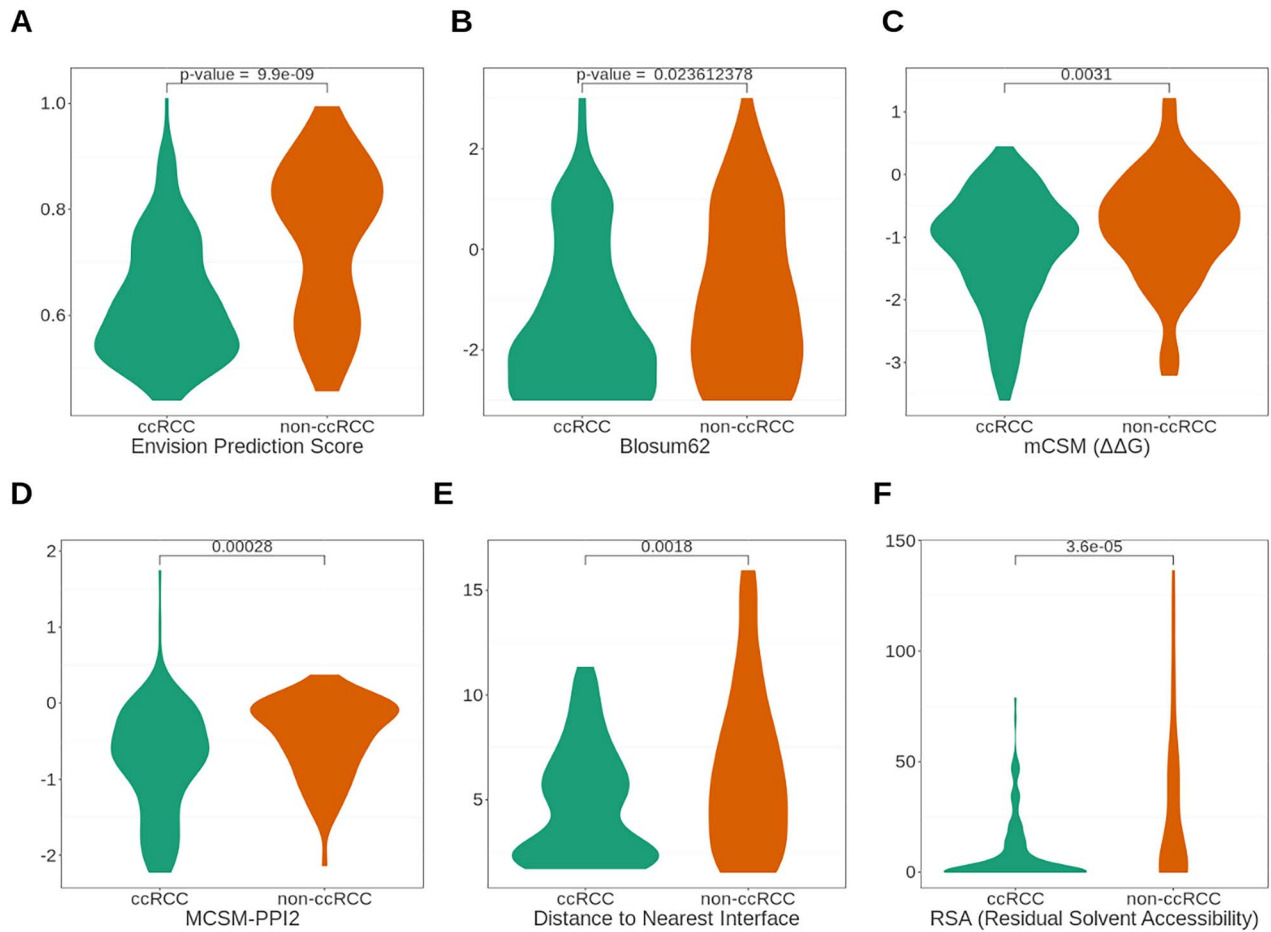


Figure 3. Comparison of key molecular features between ccRCC and non-ccRCC-causing mutations. Statistical significance was calculated using the Welch for the Envision Prediction score (A), Blosum62 (B), mCSM $\Delta\Delta G$ (C), mCSM-PPI2 (D), distance to nearest interface (E), and RSA (F).

Additionally, the thermodynamic properties of a molecule have a significant influence on its behavioral properties. Protein stability, particularly in the structured regions of a protein, is typically necessary for it to carry out its normal biological function, and the greater destabilizing effect of a mutation the greater disturbances to its normal physiological function are expected. Several tools were used to predict the change in enthalpy of the mutated protein compared to that of the wild type (e.g. DynaMut2, mCSM-Stability). Using mCSM-Stability, it was observed that there was a significant difference in predicted Gibbs free energy change between ccRCC causing and non-ccRCC causing mutations (Fig. 3C), P -value = 3.10×10^{-3} , as shown previously, with ccRCC-causing mutations leading to larger destabilizing effects [52].

Often, cancerous mutations exert their pathogenic effects by disturbing critical protein-protein interactions necessary for normal physiological function. It is for this reason, we investigated the change in protein affinity for ccRCC and non-ccRCC mutations for all proteins within the VHL complex. This was accomplished using mCSM-PPI2, a computational tool used to predict the change in protein-protein affinity from missense mutations. Using the results from mCSM-PPI2, a larger decrease in protein affinity could be seen in ccRCC-causing mutations than non-ccRCC-causing mutations (Fig. 3D), P -value = 2.80×10^{-4} .

Moreover, the spatial arrangement and packaging of the amino acid residues within a protein has enormous ramifications on its structure and function. Thus, the relative location of the amino

acid change within the VHL protein complex was investigated. Analysis of the distance of a missense mutation to the various interfaces in the VHL protein complex showed a tendency of cancer-causing mutations to be located in closer proximity to any of the protein interfaces within the VHL complex (Fig. 3E), P -value = 1.80×10^{-3} .

Another influential attribute of protein function is the solvent exposure of amino acid residue in a protein. Solvent exposure is an important factor in protein-protein interactions, recognition of protein fold and identification [53, 54]. Thus, examining the solvent exposure of different disease phenotypes may provide insight into the patterns of ccRCC mutations. Several different measures of solvent exposure are widely used, but the one of particular interest is the relative solvent accessibility (RSA), which is the distance between residue and closest surface water molecule. Examination of the RSA for residues within the VHL complex revealed ccRCC-causing mutations are localized closer within the protein core, while non-ccRCC mutations localized more toward the surface (Fig. 3F), P -value = 3.60×10^{-5} .

A new machine learning model to predict ccRCC-causing mutations

During model development, where we tested 11 different algorithms, Random Forest consistently outperformed alternative approaches. This was hence chosen as our final model, which showed a Matthew's Correlation Coefficient (MCC) value of 0.44 under 10-fold cross-validation and on the blind test (Table 2),

Table 2. Performance matrices of random forest model for 10-fold CV and blind test set.

Validation method	F1-score	Accuracy	AUC	MCC
10-fold CV	0.72	0.73	0.71	0.44
blind-test	0.80	0.81	0.77	0.44

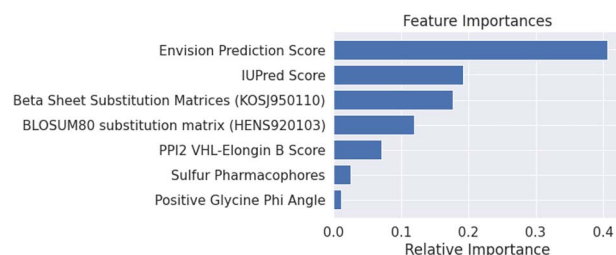
both of which were composed of data from clinical studies. This demonstrates the superior performance of our model compared to that of previous methods. Considering other metrics (Table 2) we also observed an AUC (Area Under Precision Recall Curve) of 0.71 and 0.77 for the cross validation and blind test, respectively, and accuracy scores of 0.73 (cross validation) and 0.81 (blind test). This demonstrates the robustness of our model. Using this model, predictions were calculated for each possible missense mutation within the VHL experimental crystallographic structure (Supplementary Table 2). A total of 1306 ccRCC-causing mutations and 1544 non-ccRCC-causing mutations were predicted, with no clear pattern of distribution between the two phenotypes along the 2D sequence of the VHL protein.

Discussion

VHL is a genetic disease that predisposes an individual to cysts and tumor formations within the retina, central nervous system, adrenal glands and kidneys. It is caused by the loss of both functional VHL alleles within a single cell, most often a result from missense mutations. One of the most common tumor types caused by VHL disease is ccRCC. Knowledge of the likelihood of a VHL mutation resulting in ccRCC development can aid clinicians in developing more effective screening strategies. Previous work attempted to resolve this issue by developing a machine learning model, named Symphony, to predict the development of ccRCC from missense mutations. However the performance of this model was constrained by the limited amount of clinical data and computational tools available at its time of development. It is for this reason, this study aimed to produce an up-to-date ccRCC development predictor for missense mutations in the VHL gene. This was achieved by curating the largest curated VHL missense mutation data set to date.

Data regarding ccRCC and non-ccRCC causing mutations was collected from the Symphony database, as well as a literature review of clinical studies. While these were used as the primary source of data in this study, other cancer-dedicated databases, like TCGA and COSMIC are also publicly available, and present clinically relevant cancer data to varying degrees. Because of the data heterogeneity these databases may have added to our curated set, we have chosen to forgo these sources for this current work. This ensured that our work, which focused on ccRCC mutations in VHL, was as clinically relevant as possible. Considering the multiple cancers VHL is involved in, however, we do not exclude the utility of these other databases in future expansions of this work. Predictably, analysis of said curated data revealed mutations concentrated in structured regions of pVHL30. However, no clear pattern distinguishing our classes could be observed when assessing mutation distribution, where mutations from both phenotypes were observed to overlap across pVHL30 domains. Meanwhile, the MTR results indicated that the alpha domain of the protein is more tolerant to mutations than the beta domain.

The biophysical attributes of VHL mutations were then examined, in order to provide insight into the molecular characteristics

**Figure 4.** Feature importance of Random Forest model. Examination of the relative importance of each feature in the Random Forest model, reveals that the Envision prediction score to be the most significant feature in the models prediction calculations by a substantial margin.

that contribute to tumor formation. For this purpose, statistical analysis of VHL mutations was performed using molecular features generated from the experimental crystallographic structure of VHL. This revealed multiple intriguing differences between ccRCC and non-ccRCC mutations. One such difference is the Envision prediction score, which quantifies the deleteriousness of a mutation. This score was shown to be a powerful predictor of ccRCC development. In addition to showing the highest statistically significant average difference between ccRCC and non-ccRCC mutations among all the different molecular features investigated, the Envision prediction score also possessed the highest relative importance value (Fig. 4) in the Random Forest machine learning model generated. This demonstrates the utility of the Envision score in predicting the impact of missense mutations.

In our analysis, we observed that ccRCC-causing mutations not only correlated to protein destabilization, but also showed lower amino acid conservation scores in comparison to non-ccRCC mutations. Additionally, the location of the mutated amino acid residues contributed to the likelihood of ccRCC development, as ccRCC-causing mutations were more likely to be located in buried protein regions, as well as being positioned closer to the interface with other proteins in the VHL complex.

Previous attempts have been made to predict ccRCC mutations, which integrated machine learning tools to predict ccRCC causing mutations. However, the Symphony model is limited by a small dataset and a reduced number of tools for molecular feature generation. This study overcomes these previous limitations by constructing a machine learning model, utilizing additional clinical data and current computational tools, as well as using an independent clinical testing dataset for evaluation purposes.

The model developed in this study integrates data regarding the biophysical, spatial and evolutionary differences between the wild type and mutated residues. When compared using the same dataset, our model outperformed all alternative methods, including Polyphen-2 and SIFT, which are widely used clinically to prioritize patients according to their mutation profiles.

Despite its high performance, our model also contains limitations. Due to the limitations of the crystallographic structure used for the VHL complex, the model can only generate missense mutation predictions for 150 of 213 amino acid residues in pVHL30. However, the residues that are beyond the scope of the model's prediction are located within unstructured regions of the protein and thus are unlikely to produce pathological consequences upon mutation. In fact, there has been very few published instances of ccRCC-causing mutations occurring within these unstructured regions.

A limitation to the performance of the model is the absence of p53 interaction data used to train the model. It is well established

that VHL is involved in the regulation p53, a tumor suppressor protein with a significant role in cancer prevention and development in the body. Due to lack of structural data regarding the interaction with VHL and p53, this information was not integrated into the training of the model. It is likely that with this data, the model would have greater accuracy and prediction capabilities, particularly because of the strong connection between both these genes and clinical cancer manifestation.

The loss of both functional VHL alleles within a single cell predisposes an individual to tumor formation occurring from said cell, with one of the most common types of tumor developing being ccRCC. Investigation into the biophysical, structural and evolutionary conservation characteristics of VHL mutations, showed statistically significant disparities between mutations causing ccRCC and those that do not. Furthermore, a machine learning model was constructed to accurately predict the development of ccRCC from a missense mutation within the VHL protein, which considerably outperformed previous VHL-specific and general disease classifiers.

Materials and Methods

Data collection

The primary objective of this study was to predict the physiological consequences of a missense mutation within the VHL gene. For this purpose, VHL missense mutation data was collected from the Symphony database, consisting of ccRCC-causing and non-ccRCC-causing mutations, of which, the latter may include the clinical manifestation of other cancers like hemangioblastoma and paraganglioma. Additionally, mutational ccRCC VHL data was gathered from the ClinVar database [55] (accessed August 2022) (Supplementary Table 1). The clinical involvement and labeled phenotype of mutations listed in ClinVar were verified by the literature. Only data produced from VHL patients were retained for analysis and model generation. Moreover, mutations not present in the ClinVar database but observed in patients within the clinical literature were also retained. In addition, a general literature review of clinical studies involving VHL patients was also carried out using key-word searches in the PubMed database. This data was then divided and used for the training and subsequent testing of the machine learning models (Supplementary Table 3).

Feature generation

There are numerous computational methodologies that aim at describing different aspects of effects of mutations on protein structure, function and interactions, ranging from molecular interactions, general physicochemical properties and 3-dimensional shape, to protein-level stability changes and effects on interactions with other molecules [56–62]. We used the aforementioned descriptors to represent missense mutations within the VHL gene and protein, for the purposes of generating molecular features. This is based on the premise that the degree to which mutation interatomic interactions deviate from wild type can be indicative of the pathogenicity mechanism of the mutated protein.

Molecular features were calculated using an experimental crystallographic structure describing VHL in a protein complex with Elongin B, Elongin C and HIF- α (PDB ID: 1LM8). Preprocessing was performed on the structure preceding feature generation using the software tool named Maestro, and involved the removal of water molecules and metal atoms, correction of missing residue atoms and addition of missing residues present in pVHL30.

This study implements the use of multiple *in silico* methods in order to generate numerical representations of the pVHL30 from each possible VHL missense mutation. A total of 314 features were generated (Supplementary Table 4), which can be broadly grouped into 6 different categories:

- 1) Graph-based signatures: In which the protein residue environment was represented by a graph, to model atoms and physicochemical interactions that are in close proximity [58].
- 2) Effects on thermodynamics: DUET [57], mCSM-Stability [58] and Dynamut2 [59, 61] were used to predict changes in protein stability as a change in Gibbs free energy change ($\Delta\Delta G$, in Kcal/mol).
- 3) Structural properties: several measures of the amino acids position relative to the VHL protein complex were calculated. These included Residue Depth (the average distance of the atoms of a residue from the solvent accessible surface) and the Residue solvent accessibility, or RSA (a measure of how exposed or buried a residue is in the 3D structure of a protein) generated using BioPython, distance to the protein interface, and backbone phi and psi angles.
- 4) Affinity changes: predictions generated using machine learning tools, mCSM-PPI and mCSM-PPI2 [60], regarding the effect of mutations on binding affinity between VHL and its respective interacting proteins HIF-1 α , Elongin B and Elongin C. In each case, values for mutations lying beyond 10 Å of the protein-protein interface were masked to solely retain direct effects on affinity.
- 5) Interatomic Interactions: Arpeggio was used to calculate several different subtypes of wild type interatomic interactions in a protein, as well as their changes upon mutation [56].
- 6) Sequenced-based features: The AAindex database, which contains physico-chemical and biochemical properties of amino acids, was used to produce molecular features of each substitution mutation [63]. Additionally, the BLOSUMs and PAMs substitution matrices, describing the likelihood of an amino acid change, were used to generate molecular features of physicochemical and evolutionary characteristics [64–67].

Statistical analysis

Statistical analysis was performed on all the molecular features generated to investigate the differences between the ccRCC and non-ccRCC causing mutations. Using the Welch Two Sample t-test, all the values of each feature were analyzed to determine any significant differences in mean between the cancer and non-cancer causing mutations (Supplementary Table 5). The Welch Two Sample t-test was chosen due to the large imbalance between the number of benign and pathogenic mutations in our dataset, as it does not assume that the groups have equal variances and performs well with unequal sample sizes. This analysis was carried out using the *ggsignif* software package for the R programming language. The cut-off for statistical significance of mean difference between the two populations was set to P -value < 0.05 .

Feature selection

As part of the machine learning pipeline, greedy feature selection was used in order to decrease feature redundancy, complexity and risk of overfitting in the final model. The greedy feature selection

method consisted of initially evaluating each feature independently in their capability for phenotype prediction, determined by the MCC. The MCC was chosen as a performance metric for model evaluation as it accounts for imbalances between classes in our dataset. MCC scores were generated for 10-fold cross validation, and our non-redundant test set. The best performing feature across both datasets is then selected, and combined with the remaining individually, in an iterative manner. This produces a cumulative list of feature combinations with their corresponding MCC values. This ultimately resulted in a greater performing model consisting of a lower number of features. This method was carried out using the *sci-kit learn* v.1.1.2 Python Package.

Machine learning

Model generation was performed using *sci-kit learn* v.1.1.2, based on the training dataset, while a non-redundant test set was used to assess model performance and generalization.

The training dataset consisted of data obtained from the Symphony database ($n = 121$; ccRCC:67; non-ccRCC:54) and the non-redundant test set was manually curated from the literature ($n = 92$; ccRCC:75; non-ccRCC: 17). This was performed in order to produce a model directly comparable to the Symphony predictor, especially when subjected to an external blind test. The predictive power of the resulting model was assessed by calculating MCC on the test set. This evaluation metric was used due to an unbalanced test dataset, a result of publication bias in the clinical literature, as the likelihood of the publication of a pathogenic variant is far greater than that of a benign one.

During model development, several machine learning algorithms were assessed in their predictive capabilities, including Gradient Boosting, XGBoost, Random Forest, Extra Trees, Gaussian Process, AdaBoost, K-nearest neighbors, Support Vector Machine, Multi-layer Perceptron, Decision Tree, and Logistic Regression algorithms. The resultant models were assessed using MCC under 10-fold cross validation. In order to increase the clinical applicability of our work, the final model was used to predict the ccRCC-risk of each possible VHL missense through an *in silico* saturation mutagenesis approach.

Author contributions

A.S. performed all data curation, machine learning and analysis, and wrote the first draft of the manuscript. G.T. and C.S. assisted with data curation and initial analysis. Q.P. assisted with feature engineering. D.E.V.P. provided guidance on data curation. S.P. helped supervise the project. D.B.A. designed and supervised all aspects of the project. All authors reviewed the manuscript.

Supplementary data

[Supplementary data](#) is available at *HMG Journal* online.

Conflict of interest statement: The authors declare no conflict of interests.

Funding

This work was supported by an Investigator Grant from the National Health and Medical Research Council (NHMRC) of Australia (GNT1174405 to D.B.A.); Supported in part by the Victorian Government's Operational Infrastructure Support Program.

Data availability

All data curated and generated in this study is available in the supplementary tables.

References

- Iliopoulos O, Kaelin WG Jr. The molecular basis of von Hippel-Lindau disease. *Mol Med* 1997;**3**:289–93.
- Kaelin WG Jr, Maher ER. The VHL tumour-suppressor gene paradigm. *Trends Genet* 1998;**14**:423–6.
- Chittiboina P, Lonser RR. Von Hippel-Lindau disease. *Handb Clin Neurol* 2015;**132**:139–56.
- Maher ER, Neumann HP, Richard S. von Hippel-Lindau disease: a clinical and scientific review. *Eur J Hum Genet* 2011;**19**:617–23.
- Nordstrom-O'Brien M, van der Luijt RB, van Rooijen E. et al. Genetic analysis of von Hippel-Lindau disease. *Hum Mutat* 2010;**31**:521–37.
- Latif F, Tory K, Gnarr J. et al. Identification of the von Hippel-Lindau disease tumor suppressor gene. *Science* 1993;**260**:1317–20.
- Gossage L, Eisen T, Maher ER. VHL, the story of a tumour suppressor gene. *Nat Rev Cancer* 2015;**15**:55–64.
- Duan DR, Pause A, Burgess WH. et al. Inhibition of transcription elongation by the VHL tumor suppressor protein. *Science* 1995;**269**:1402–6.
- Kibel A, Iliopoulos O, DeCaprio JA. et al. Binding of the von Hippel-Lindau tumor suppressor protein to Elongin B and C. *Science* 1995;**269**:1444–6.
- Kishida T, Stackhouse TM, Chen F. et al. Cellular proteins that bind the von Hippel-Lindau disease gene product: mapping of binding domains and the effect of missense mutations. *Cancer Res* 1995;**55**:4544–8.
- Loneragan KM, Iliopoulos O, Ohh M. et al. Regulation of hypoxia-inducible mRNAs by the von Hippel-Lindau tumor suppressor protein requires binding to complexes containing elongins B/C and Cul2. *Mol Cell Biol* 1998;**18**:732–41.
- Stebbins CE, Kaelin WG Jr, Pavletich NP. Structure of the VHL-ElonginC-ElonginB complex: implications for VHL tumor suppressor function. *Science* 1999;**284**:455–61.
- Haase VH. The VHL/HIF oxygen-sensing pathway and its relevance to kidney disease. *Kidney Int* 2006;**69**:1302–7.
- Kaelin WG. The von Hippel-Lindau tumor suppressor protein: roles in cancer and oxygen sensing. *Cold Spring Harb Symp Quant Biol* 2005;**70**:159–66.
- Kaelin WG. von Hippel-Lindau disease. *Annu Rev Pathol* 2007;**2**:145–73.
- Shen C, Kaelin WG Jr. The VHL/HIF axis in clear cell renal carcinoma. *Semin Cancer Biol* 2013;**23**:18–25.
- Roe JS, Youn HD. The positive regulation of p53 by the tumor suppressor VHL. *Cell Cycle* 2006;**5**:2054–6.
- Semenza GL. VHL and p53: tumor suppressors team up to prevent cancer. *Mol Cell* 2006;**22**:437–9.
- Tichkule S, Myung Y, Naung MT. et al. VIVID: a web application for variant interpretation and visualization in multi-dimensional analyses. *Mol Biol Evol* 2022;**39**:msac196, p.1–8.
- Pandurangan AP, Ascher DB, Thomas SE. et al. Genomes, structural biology and drug discovery: combating the impacts of mutations in genetic disease and antibiotic resistance. *Biochem Soc Trans* 2017;**45**:303–11.
- Airey E, Portelli S, Xavier JS. et al. Identifying genotype-phenotype correlations via integrative mutation analysis. *Methods Mol Biol* 2021;**2190**:1–32.

22. Pires DE, Chen J, Blundell TL. *et al.* In silico functional dissection of saturation mutagenesis: interpreting the relationship between phenotypes and changes in protein stability, interactions and activity. *Sci Rep* 2016;**6**:19848.
23. Portelli S, Albanaz A, Pires DEV. *et al.* Identifying the molecular drivers of ALS-implicated missense mutations. *J Med Genet* 2022;**60**:484–90.
24. Stephenson SEM, Costain G, Blok LER. *et al.* Germline variants in tumor suppressor FBXW7 lead to impaired ubiquitination and a neurodevelopmental syndrome. *Am J Hum Genet* 2022;**109**:601–17.
25. Karmakar M, Cicaloni V, Rodrigues CHM. *et al.* HGDiscovery: an online tool providing functional and phenotypic information on novel variants of homogenisate 1,2-dioxygenase. *Curr Res Struct Biol* 2022;**4**:271–7.
26. Lai CY, Tsai IJ, Chiu PC. *et al.* A novel deep intronic variant strongly associates with Alkaptonuria. *NPJ Genom Med* 2021;**6**:89.
27. Hildebrand JM, Kauppi M, Majewski IJ. *et al.* A missense mutation in the MLKL brace region promotes lethal neonatal inflammation and hematopoietic dysfunction. *Nat Commun* 2020;**11**:3150.
28. Ascher DB, Spiga O, Sekelska M. *et al.* Homogenisate 1,2-dioxygenase (HGD) gene variants, their analysis and genotype-phenotype correlations in the largest cohort of patients with AKU. *Eur J Hum Genet* 2019;**27**:888–902.
29. Parthasarathy S, Ruggiero SM, Gelot A. *et al.* A recurrent de novo splice site variant involving DNMT1 exon 10a causes developmental and epileptic encephalopathy through a dominant-negative mechanism. *Am J Hum Genet* 2022;**109**:2253–69.
30. Soardi FC, Machado-Silva A, Linhares ND. *et al.* Familial STAG2 germline mutation defines a new human cohesinopathy. *NPJ Genom Med* 2017;**2**:7.
31. Portelli S, Barr L, de Sa AGC. *et al.* Distinguishing between PTEN clinical phenotypes through mutation analysis. *Comput Struct Biotechnol J* 2021;**19**:3097–109.
32. Aljarf R, Shen M, Pires DEV. *et al.* Understanding and predicting the functional consequences of missense mutations in BRCA1 and BRCA2. *Sci Rep* 2022;**12**:10458.
33. Andrews KA, Ascher DB, Pires DEV. *et al.* Tumour risks and genotype-phenotype correlations associated with germline variants in succinate dehydrogenase subunit genes SDHB, SDHC and SDHD. *J Med Genet* 2018;**55**:384–94.
34. Casey RT, Ascher DB, Rattenberry E. *et al.* SDHA related tumorigenesis: a new case series and literature review for variant interpretation and pathogenicity. *Mol Genet Genomic Med* 2017;**5**:237–50.
35. Hnizda A, Fabry M, Moriyama T. *et al.* Relapsed acute lymphoblastic leukemia-specific mutations in NT5C2 cluster into hotspots driving intersubunit stimulation. *Leukemia* 2018;**32**:1393–403.
36. Jafri M, Wake NC, Ascher DB. *et al.* Germline mutations in the CDKN2B tumor suppressor gene predispose to renal cell carcinoma. *Cancer Discov* 2015;**5**:723–9.
37. Zhou Y, Portelli S, Pat M. *et al.* Structure-guided machine learning prediction of drug resistance mutations in Abelson 1 kinase. *Comput Struct Biotechnol J* 2021;**19**:5381–91.
38. Karmakar M, Globan M, Fyfe JAM. *et al.* Analysis of a novel pncA mutation for susceptibility to pyrazinamide therapy. *Am J Respir Crit Care Med* 2018;**198**:541–4.
39. Karmakar M, Rodrigues CHM, Holt KE. *et al.* Empirical ways to identify novel bedaquiline resistance mutations in AtpE. *PLoS One* 2019;**14**:e0217169.
40. Karmakar M, Rodrigues CHM, Horan K. *et al.* Structure guided prediction of pyrazinamide resistance mutations in pncA. *Sci Rep* 2020;**10**:1875.
41. Portelli S, Myung Y, Furnham N. *et al.* Prediction of rifampicin resistance beyond the RRDR using structure-based machine learning approaches. *Sci Rep* 2020;**10**:18120.
42. Portelli S, Olshansky M, Rodrigues CHM. *et al.* Exploring the structural distribution of genetic variation in SARS-CoV-2 with the COVID-3D online resource. *Nat Genet* 2020;**52**:999–1001.
43. Portelli S, Phelan JE, Ascher DB. *et al.* Understanding molecular consequences of putative drug resistant mutations in *Mycobacterium tuberculosis*. *Sci Rep* 2018;**8**:15356.
44. Vediti SC, Malhotra S, Skwark MJ. *et al.* HARP: a database of structural impacts of systematic missense mutations in drug targets of *Mycobacterium leprae*. *Comput Struct Biotechnol J* 2020;**18**:3692–704.
45. Hawkey J, Ascher DB, Judd LM. *et al.* Evolution of carbapenem resistance in *Acinetobacter baumannii* during a prolonged infection. *Microb Genom* 2018;**4**:e000165, p. 1–11. <https://doi.org/10.1099/mgen.0.000165>.
46. Silk M, Petrovski S, Ascher DB. MTR-viewer: identifying regions within genes under purifying selection. *Nucleic Acids Res* 2019;**47**:W121–6.
47. Silk M, Pires DEV, Rodrigues CHM. *et al.* MTR3D: identifying regions within protein tertiary structures under purifying selection. *Nucleic Acids Res* 2021;**49**:W438–45.
48. Traynelis J, Silk M, Wang Q. *et al.* Optimizing genomic medicine in epilepsy through a gene-customized approach to missense variant interpretation. *Genome Res* 2017;**27**:1715–29.
49. Adzhubei IA, Schmidt S, Peshkin L. *et al.* A method and server for predicting damaging missense mutations. *Nat Methods* 2010;**7**:248–9.
50. Ng PC, Henikoff S. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res* 2003;**31**:3812–4.
51. Gray VE, Hause RJ, Luebeck J. *et al.* Quantitative missense variant effect prediction using large-scale mutagenesis data. *Cell Syst* 2018;**6**:116–124.e3.
52. Gossage L, Pires DE, Olivera-Nappa A. *et al.* An integrated computational approach can classify VHL missense mutations according to risk of clear cell renal carcinoma. *Hum Mol Genet* 2014;**23**:5976–88.
53. Rodrigues CHM, Pires DEV, Blundell TL. *et al.* Structural landscapes of PPI interfaces. *Brief Bioinform* 2022;**23**:bbac165, p. 1–10.
54. Jubb HC, Pandurangan AP, Turner MA. *et al.* Mutations at protein-protein interfaces: small changes over big surfaces have large impacts on human health. *Prog Biophys Mol Biol* 2017;**128**:3–13.
55. Landrum MJ, Lee JM, Riley GR. *et al.* ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res* 2014;**42**:D980–5.
56. Jubb HC, Higuero AP, Ochoa-Montano B. *et al.* Arpeggio: a web server for calculating and visualising interatomic interactions in protein structures. *J Mol Biol* 2017;**429**:365–71.
57. Pires DE, Ascher DB, Blundell TL. DUET: a server for predicting effects of mutations on protein stability using an integrated computational approach. *Nucleic Acids Res* 2014;**42**:W314–9.
58. Pires DE, Ascher DB, Blundell TL. mCSM: predicting the effects of mutations in proteins using graph-based signatures. *Bioinformatics* 2014;**30**:335–42.

59. Rodrigues CH, Pires DE, Ascher DB. DynaMut: predicting the impact of mutations on protein conformation, flexibility and stability. *Nucleic Acids Res* 2018;**46**:W350–5.
60. Rodrigues CHM, Myung Y, Pires DEV. et al. mCSM-PPI2: predicting the effects of mutations on protein-protein interactions. *Nucleic Acids Res* 2019;**47**:W338–44.
61. Rodrigues CHM, Pires DEV, Ascher DB. DynaMut2: assessing changes in stability and flexibility upon single and multiple point missense mutations. *Protein Sci* 2021;**30**:60–9.
62. Pandurangan AP, Ochoa-Montano B, Ascher DB. et al. SDM: a server for predicting effects of mutations on protein stability. *Nucleic Acids Res* 2017;**45**:W229–35.
63. Kawashima S, Kanehisa M. AAindex: amino acid index database. *Nucleic Acids Res* 2000;**28**:374.
64. Fabris F, Sgarro A, Tossi A. Splitting the BLOSUM score into numbers of biological significance. *EURASIP J Bioinform Syst Biol* 2007;**2007**:31450.
65. Mihalek I, Res I, Lichtarge O. Background frequencies for residue variability estimates: BLOSUM revisited. *BMC Bioinformatics* 2007;**8**:488.
66. Mount DW. Using PAM matrices in sequence alignments. *CSH Protoc* 2008;**2008**:pdb.top38.
67. Muller T, Vingron M. Modeling amino acid replacement. *J Comput Biol* 2000;**7**:761–76.