Computer Methods and Programs in Biomedicine 219 (2022) 106768

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

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

Unveiling the influence of factor VIII physicochemical properties on hemophilia A phenotype through an *in silico* methodology

Mariana R. Meireles[a], Lara H. Stelmach[a], Eliane Bandinelli[a], Gustavo F. Vieira[a,b,*]

[a] Departamento de Genética, Instituto de Biociências, Universidade Federal do Rio Grande do Sul, Caixa Postal 15053, Porto Alegre 91501-970, RS, Brasil
[b] Universidade La Salle, Canoas, RS, Brasil

ARTICLE INFO

Article history:
Received 21 June 2021
Revised 24 January 2022
Accepted 19 March 2022

Keywords:
Missense mutation
Coagulation
Molecular properties
Bioinformatics
Functional impact
Mendelian disorders

ABSTRACT

Background and objectives: Hemophilia A (HA) is an X-linked blood disorder. It is caused by pathogenic *F8* gene variants, among which missense mutations are the most prevalent. The resulting amino acid substitutions may have different impacts on physicochemical properties and, consequently, on protein functionality. Regular prediction tools do not include structural elements and their physiological significance, which hampers our ability to functionally link variants to disease phenotype, opening an ample field for investigation. The present study aims to elucidate how physicochemical changes generated by substitutions in different protein domains relate to HA, and which of these features are more consequential to protein function and its impact on HA phenotype.

Methods: An *in silico* evaluation of 71 *F8* variants found in patients with different HA phenotypes (mild, moderate, severe) was performed to understand protein modifications and functional impact. Homology modeling was used for the structural analysis of physicochemical changes including electrostatic potential, hydrophobicity, solvent-accessible/excluded surface areas, disulfide disruptions, and substitutions indexes. These variants and properties were analyzed by hierarchical clustering analysis (HCA) and principal component analysis (PCA), independently and in combination, to investigate their relative contribution.

Results: About 69% of variants show electrostatic changes, and almost all show hydrophobicity and surface area modifications. HCA combining all physicochemical properties analyzed was better in reflecting the impact of different variants in disease severity, more so than the single feature analysis. On the other hand, PCA led to the identification of prominent properties involved in the clustering results for variants of different domains.

Conclusions: The methodology developed here enables the assessment of structural features not available in other prediction tools (e.g., surface distribution of electrostatic potential), evaluating what kind of physicochemical changes are involved in FVIII functional disruption. HCA results allow distinguishing substitutions according to their properties, and yielded clusters which were more homogeneous in phenotype. All evaluated properties are involved in determining disease severity. The nature, as well as the position of the variants in the protein, were shown to be relevant for physicochemical changes, demonstrating that all these aspects must be collectively considered to fine-tune an approach to predict HA severity.

1. Introduction

Hemophilia A (HA) is the most prevalent X-linked bleeding disorder affecting about 1: 5000 male live births [1]. The condition is caused by mutations in the *F8* gene which encodes Factor VIII (FVIII), a crucial protein in the coagulation process [2]. Mature FVIII is composed of six domains: three A domains (A1, A2, and A3), one B domain, and two C domains (C1 and C2) [3,4]. Proteolytic processing by thrombin cleaves several Arginine residues, leading to the removal of the B domain and secretion of FVIII to the plasma [5,6]. According to the International Society of Thrombosis and Haemostasis, HA is classified based on the plasmatic level of FVIII as mild (5–40%), moderate (1–5%), and severe (<1%) [7].

The disease manifestations involve a broad clinical heterogeneity with a range of causative variants. One deleterious variant in the *F8* gene is sufficient to cause hemophilia. According to the European Association of Hemophilia and Allied Dis-

* Corresponding author at: Programa de Pós-graduação em Saúde e Desenvolvimento Humano. Av. Victor Barreto, 2288, Centro - Canoas RS, 92010-000.
E-mail address: gustavo.vieira@unilasalle.edu.br (G.F. Vieira).

https://doi.org/10.1016/j.cmpb.2022.106768
0169-2607/

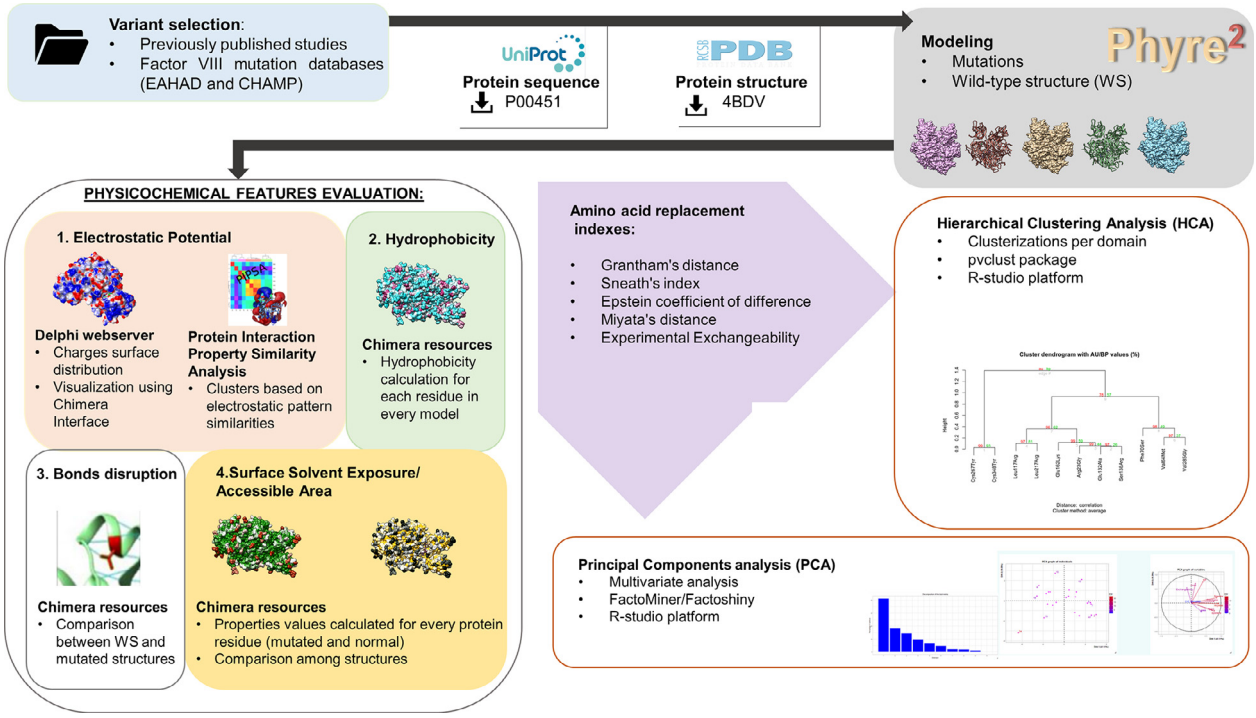

Fig. 1. *Methodological pipeline*. **A-** Selecting pathogenic missense variants based on previous descriptions in the European Association for Hemophilia and Allied Disorders Database (EAHAD) and reported studies. **B**- Generating models using the Phyre-2 software in its expert mode "one-to-one threading job" based on the PDB template 4BDV and UniProt FASTA sequence P00451. **C**- Evaluating physicochemical properties of the models: (1) Electrostatic surface pattern and EP similarities were obtained using the Delphi web server and PIPSA software, respectively; Application of Chimera interface resources examining (2) hydrophobicity, (3) bonds disruptions, and (3) solvent interactions surface areas. D- Calculating amino acids replacement indexes calculations. **E-** Hierarchical clustering analysis (HCA) combining the structure features (C-1,2,3 and 4) and indexes (D) through the pvclust package (R-Studio platform). And **F-** a Principal Component Analysis (PCA) using the FactoMineR package (R-Studio).

orders (EAHAD), 66.2% of *F8* variations are point mutations, with missense being the most prevalent of them (more than 70%) [8,9]. Evaluating the impact of amino acid substitutions is considered a challenge since these variations occur in a spectrum that includes benign to pathogenic changes. Many factors related to amino acid substitution could impact FVIII functionality, and consequently, disease severity: amino acid properties and position, bond disruptions, modifications that affect protein conformation and affinity with coagulation cascade partners.

The present work aims to evaluate how physicochemical properties affect FVIII functionality, analyzing the impact of specific residue changes to improve our understanding of genotype-phenotype correlations in HA. The study calculates and examines physicochemical features including electrostatic potential (EP), hydrophilic/hydrophobic changes, hydrogen and disulfide bond alterations, solvent-accessible and excluded surface areas. Combined distance indexes of substitutions were applied improving the analysis. All cited features were applied to understand how missense variants affect protein functionality and the consequent effect in phenotype, identifying the principal components involved in alterations.

2. Methods

Fig. 1 summarizes the methodological pipeline. The accession numbers of NCBI sequences used are: NG_011403.1 (*F8*) for genome and NM_000132.4 (*F8*) for the transcript, and the UniProt accession code is P00451. Variant nomenclature was verified using the Name Checker tool from Mutalyzer (https://mutalyzer.nl/name-checker), as is presented in **Supplemental Material.**

2.1. Selection of variants

We selected 71 different pathogenic variants located across the five mature FVIII domains (A1, A2, A3, C1, and C2). All the selected variants were previously described as causative of hemophilia A in worldwide publications and/or in hemophilia databases from the Centers for Disease Control and Prevention (CDC) and European Association for Hemophilia and Allied Disorders (EAHAD), with specified phenotype according to plasmatic level of FVIII. All selected variants and their information are presented in Table 1. Among the71 chosen substitutions, 26 occur in mild, 10 in moderate, and 35 in severe HA phenotype (approximately 37%, 14% and 49%, respectively). To differentiate effects related to the position rather than the nature of substituted residues, we selected multiple substitutions occurring at given positions (2 or 3 substitutions for 12 of the sites, totalling 27 variants. We prioritized variants associated with mild and severe phenotypes to highlight more categorical changes in protein activity. The number of analyzed variants allows for some generalizations to be extracted for future use in developing a broader tool.

2.2. Modeling approach

The FASTA file (text-based format) containing the protein sequence from human FVIII was manually edited to obtain customized substituted sequences and the wild type (WT). Sequences were applied in a modeling approach using Phyre-2 software [10] through its expert one-to-one threading mode, which enables the use of the sequences combined with a PDB template (4BDV, Resolution: 3.98 Å) to produce structural models. The 4BDV template and P00751 (UniProt) sequence were aligned in NCBI. The regions comprising A1 and A2 domains (residues 20–730) and A3, C1

Table 1
Missense mutations included in the study and their information.

Protein †	Transcript						Protein	Transcript					
p.Arg23Gly	c.67A>G	1	A1	Mild	5–6	[1]	p.Ile494Thr	c.1481T>C	10	A2	mild	6–7	[2]
p.Glu30Val	c.89A>T	1	A1	Mild	unreported	[3]	p.Gly498Arg	c.1492G>A	10	A2	severe	<1	[4]
p.Val64Met	c.190G>A	2	A1	Mild	20	[5]	p.Tyr530Cys	c.1589A>G	11	A2	mild	8–11	[6]
p.Lys67Glu	c.199A>*G*	2	A1	Severe	<1	[7]	p.Asp544Glu	c.1632T>*A*	11	A2	severe	<1	[8]
p.Phe70Ser	c.209T>C	2	A1	Severe	<1	[9]	p.Ser553Pro	c.1657T>C	11	A2	severe	<1	[10]
p.Leu117Arg	c.350T>*G*	3	A1	Severe	<1	[11]	p.Cys547Arg	c.1639T>C	11	A2	moderate	unreported	[12]
p.Glu132Ala	c.395A>C	4	A1	Mild	5	[5]	p.Cys547Tyr	c.1640G>*A*	11	A2	severe	<1	[9]
p.Ser138Arg	c.414T>*G*	4	A1	Mild	unreported	[13]	p.Cys547Trp	c.1641C>G	11	A2	severe	<1	[14]
p.Glu162Lys	c.484G>*A*	4	A1	Mild	26	[9]	p.Ser635Asn	c.1904G>*A*	13	A2	mild	6–28	[7]
p.Gly201Glu	c.602G>*A*	5	A1	Mild	7–13	[12]	p.Cys649Arg	c.1945T>C	13	A2	severe	<1	[15]
p.Leu217Arg	c.650T>*G*	5	A1	Severe	<1	[16]	p.Cys649Ser	c.1945T>*A*	13	A2	Severe	<1	[17]
p.His228Arg	c.683A>*G*	6	A1	Mild	2–7	[12]	p.Gly705Ser	c.2113G>*A*	13	A2	severe	<1	[16]
p.Arg259Thr	c.776G>C	6	A1	Mild	14.9	[18]	p.Thr1714Ser	c.5140A>*T*	14	A3	Moderate	3	[19]
p.Cys267Tyr	c.800G>*A*	7	A1	Moderate	3	[20]	p.Arg1740Met	c.5219G>*T*	14	A3	Severe	<1	[12]
p.Val285Gly	c.855T>*G*	7	A1	Mild	unreported	[21]	p.Lys1751Arg	c.5252A>*G*	15	A3	Severe	<1	[22]
p.Cys348Tyr	c.1043G>*A*	8	A1	Severe	<1	[23]	p.Lys1751Gln	c.5251A>C	15	A3	Severe	<1	[16]
p.Arg437Pro	c.1310G>C	9	A2	Severe	<1	[13]	p.Lys1751Glu	c.5251A>*G*	15	A3	Severe	<1	[24]
p.Lys444Arg	c.1331A>*G*	9	A2	severe	<1	[25]	p.Gln1764Arg	c.5291A>*G*	15	A3	Severe	<1	[26]
p.Lys444Asn	c.1332A>*T*	9	A2	severe	<1	[27]	p.Arg1768Cys	c.5302C>T	15	A3	Mild	30	[28]
p.Tyr450Cys	c.1349A>*G*	9	A2	mild	10	[6]	p.Arg1768His	c.5303G>*A*	15	A3	Mild	37	[29]
p.Arg458Cys	c.1372C>T	9	A2	mild	17	[30]	p.Arg1783Gly	c.5347A>*G*	15	A3	Mild	unreported	[28]
p.Arg1783Thr	c.5348G>C	15	A3	Mild	52	[1]	p.Cys2040Tyr	c.6119G>*A*	20	C1	Moderate	11	[31]
p.Arg1800His	c.5399G>*A*	16	A3	Moderate	3	[28]	p.Pro2067Leu	c.6200C>T	21	C1	Severe	<1	[12]
p.Arg1800Pro	c.5399G>C	16	A3	Severe	<1	[1]	p.Ala2080Thr	c.6238G>*A*	21	C1	Moderate	5	[25]
p.Tyr1805Cys	c.5414A>*G*	16	A3	Moderate	4	[32]	p.Arg2169Cys	c.6505C>T	23	C1	Mild	19–47	[12]
p.Tyr1805Phe	c.5414A>*T*	16	A3	Mild	>5	[28]	p.Arg2169Leu	c.6506G>*T*	23	C1	Mild	19–33	[28]
p.Tyr1805Ser	c.5414A>C	16	A3	Mild	10–14	[5]	p.Cys2188Tyr	c.6563G>*A*	23	C1	Severe	<1	[10]
p.Tyr1811Cys	c.5432A>*G*	16	A3	Mild	>5	[28]	p.Cys2193Gly	c.6577T>*G*	23	C2	Moderate	5	[33]
p.Leu1875Pro	c.5624T>C	17	A3	Severe	<1	[34]	p.Pro2224Arg	c.6671C>G	24	C2	Severe	<1	[35]
p.Leu1882Pro	c.5645T>C	17	A3	Severe	<1	[31]	p.Pro2311His	c.6932C>A	26	C2	Mild	14–37	[28]
p.Arg1960Leu	c.5879G>*T*	18	A3	Moderate	7	[36]	p.Pro2311Ser	c.6931C>T	26	C2	Mild	>5	[1]
p.Arg1960Gln	c.5879G>*A*	18	A3	Mild	>5	[28]	p.Leu2316Arg	c.6947T>*G*	26	C2	Severe	<1	[38]
p.Trp1961Arg	c.5881T>*A*	18	A3	Severe	<1	[39]	p.Leu2316Pro	c.6947T>C	26	C2	Severe	<1	[39]
p.Asn1971Thr	c.5912A>C	18	A3	Severe	<1	[11]	p.Cys2345Ser	c.7034G>C	26	C2	Severe	<1	[40]
p.Gly2013Arg	c.6037G>*A*	19	A3	Severe	<1	[35]	p.Cys2345Tyr	c.7034G>*A*	26	C2	Severe	<1	[34]
p.Tyr2036Cys	c.6107A>*G*	19	A3	Moderate	3	[41]							

The variant nomenclature was verified in Name Checker service at Mutalyzer 2.0.32 (available at https://mutalyzer.nl/name-checker with NCBI accession NM_000132.4*(F8)*.

† Sequence changes at protein level follow the Nomenclature for Description of Genetic Variations approved by Human Genome Variation Society (HGVS).The variants reference list is available at Supplementary material.Most of the variants' references were extracted from the EAHAD (European Association for Hemophilia and Allied Disorders) "Factor VIII Gene (*F8)* Variant Database" or the CHAMP – "CDC Hemophilia A mutation project"- mutation list.

and C2 domains (residues 1667–2351) are 100% identical to 4BDV chain A and B, respectively (Supplemental Material).

The models comprise the five domains present in circulating activated FVIII (A1, A2, A3, C1, and C2). A total of 76 structures were generated: 17 models for A1, being 16 variants and the wild-type structure (WS), 18 for A2 (17 variants plus WS), 25 for A3 (24 variants plus WS), 7 for C1 (6 variants plus WS), and 9 for C2 (8 variants plus WS). All models were subjected to quality evaluation using ModFold [11,12]and PDBsum [13].

2.2.1. Physicochemical analyses

The models were inputted in the Delphi web server tool [14,15], which allows electrostatic potential (EP) calculations, as well as visual inspection of EP distribution through its application in the Chimera interface [16]. We also performed a clustering analysis exclusively based on EP property using the webPIPSA software [17]. This analysis considered the domains independently. The use of both procedures enables a comprehensive linking of EP modifications to disease outcomes. Altoghether, these analyses identify both the surface change alterations in each model and the electrostatic distance between structures.

Other features were assessed via Chimera resources: hydrophobicity changes, hydrogen bonds, and cysteine bond disruptions, solvent-accessible surface area (areaSAS), and solvent-excluded surface area (areaSES) modifications. The values of each attribute were extracted from models individually and compared to the respective wild-type structure.

2.3. Distance index

Scores for distances among variants according to amino acid substitutions were calculated by the following matrices: Grantham's distance [18], Sneathś index [19], Epsteinś coefficient of distance [20], Miyata's distance [21], and Experimental Exchangeability [22].

2.4. Hierarchical clustering analysis (HCA)

The evaluated physicochemical properties were collectively applied as input for Hierarchical Clustering Analysis (HCA) performed through the R Studio platform, using the pvclust package [23]. The output returns clusters and branches with associated probability values using bootstrap resampling techniques: approximately unbiased (AU), obtained with multiscale bootstrap resampling, and bootstrap probability (BP), calculated by the ordinary bootstrap resampling technique. Generally, the AU p-value is a better estimator of the reliability of the clusters obtained, and was therefore used as the statistical metric in our work.

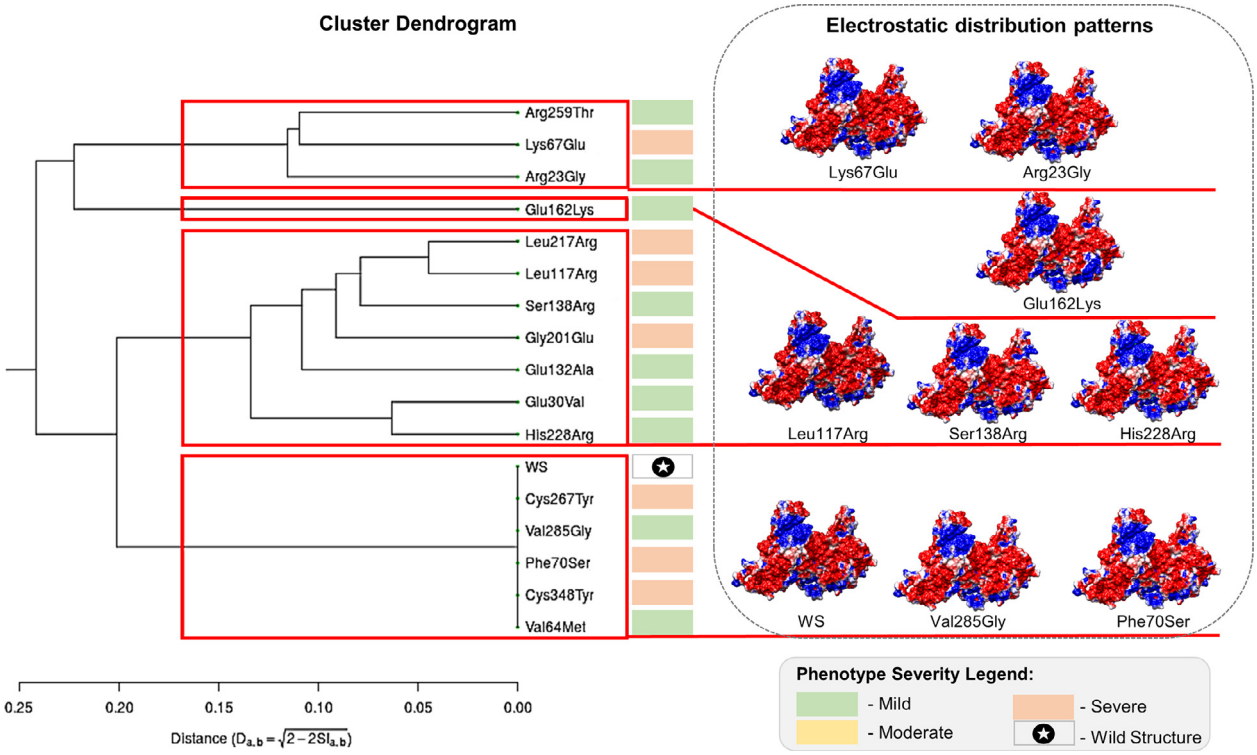

Fig. 2. *Electrostatic potential similarities among A1 domain structural models*. The webPIPSA tool group models according to their electrostatic potential (EP) similarities (left). Delphi web server calculations for EP inputted in Chimera interface allowed obtaining of models EP surface distribution (right). The phenotype associated with each model is specified by the colors legend in the box. Cluster 4 grouped the wild-type structure with five pathogenic variants exhibiting distinct HA severity (showing no distance among them), indicating the participation of other properties in phenotype determination.

2.5. Principal components analysis (PCA)

We performed a multivariate analysis using the FactoMineR package (FactoShiny) in the R cross-platform [24]. Subsequent Principal Component Analysis (PCA) indicated the most contributive attributes in a complex scenario, like the variance presented by the spectrum of variants properties.

3. Results

3.1. Physicochemical alterations

The EP dendrograms and surface distribution results are depicted in Figs. 2**, S1–S4** (for A1, A2, A3, C1, and C2 domains, respectively). Based on electrostatic features, about 69% of the variants (49/71) form clusters separated from the cluster containing wild-type structure (WS). These distance values varied according to the domain analyzed and the disease severity. Among A1 domain variants, 68.75% clustered separately from WS, while this ratio was 35.3%, 83.3%, 66.67%, and 87.5%, for A2, A3, C1, and C2, respectively. Among variants connected to mild or moderate HA, 71% were separated from WS, in contrast to 64.7% among variants of severe phenotype. Interestingly, 85,2% (23) of the 27 variants corresponding to multiple substitutions at a given site did not cluster together with their counterparts in the dendrogram. The visual inspection of EP distribution revealed distinct patterns consistent with the epigrams distances (major distance divergences are more readily recognizable at model surfaces).

Except for p.Asp544Glu (A2 domain), every substitution translated into modifications in hydrophobicity values when compared with WS (Fig. 3 for A1 domain and **Figs. S6–S9** for A2, A3, C1, and C2, respectively). While some changes are bland and thus harder to perceive, others are more evident because they shift the local hydrophobicity from negative (more hydrophilic residues) to positive or vice versa, as can be observed for p.Cys547 and p.Ile494 substitutions within the A1 domain (Fig. 3).

Disulfide-bonds disruptions occur in 12/71 variants included in the present study. Among variants of the A1 domain, two showed disulfide disruptions, corresponding to 12.5% (2/16 – p.Cys267Tyr, p.Cys348Tyr), while within the A2 domain, 29.4% of the variants presented disrupted bonds (5/17 - p.Cys547Arg, p.Cys547Tyr, p.Cys547Trp, p.Cys649Arg, p.Cys649Ser). Variants with disulfide-bond disruptions represented 28.6% (2/7- p.Cys2040Tyr, p.Cys2188Tyr) and 37.5% (3/8- p.Cys2193Gly, p.Cys2345Ser, p.Cys2345Tyr) among C1 and C2 domain variants, respectively.

The solvent-accessible and solvent-excluded surface areas (areaSAS and areaSES, respectively) are modified in most of the chosen variants, when compared with WS, except for p. Ala2080Thr at the C1 domain. Every mutated model shows modifications in one or more of the analyzed attributes, evidencing the importance of a combined analysis.

3.2. Hierarchical clustering analysis

The HCA results pertaining to domains A1, A2, A3, C1 and C2 are shown in Figs. 4**, S11, S12, S13,** and **S14**, respectively. From all 71 variants only p.Pro2311His (mild phenotype, C2 domain) was grouped with WS (Fig. S14), which means 98.6% of the variants were markedly distinct from the wild-type. About 66.7% (18/27) of the variants affecting the same residues were in distinct clusters from their counterparts (Fig. 6).

A second HCA was performed for A1 domain, using only the index values matrices (Fig. S10). Although no pathogenic variant clustered with the WS, the number of divergent intra-cluster phenotypes increased. While the multiple features HCA (Fig. 4)

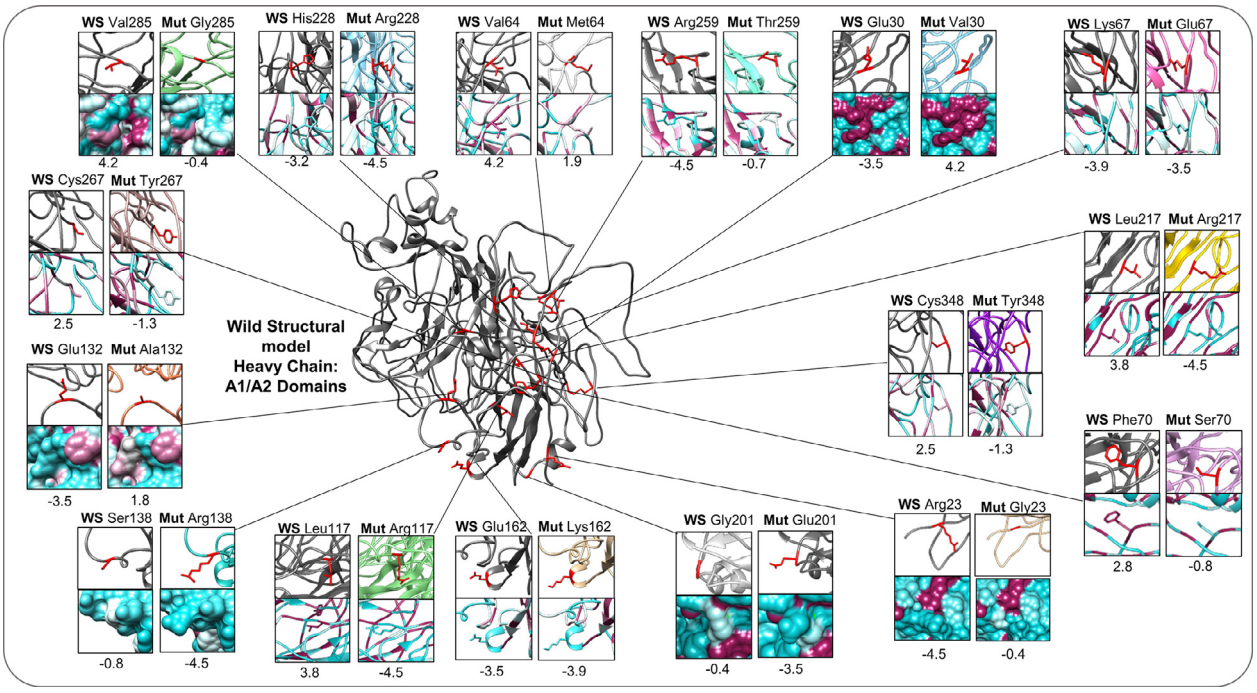

Fig. 3. *Analyzing the hydrophobicity feature in A1 domain variants*. The wild-type structure model in the center shows amino acids sites in the protein chain. The boxes depict every specific variant change in hydrophobicity considering the mutated residues (Mut) in comparison to wild type (WS). The hydrophobicity scale values vary from –4.5 (more hydrophilic residues) to 4.5 (more hydrophobic ones). The maroon pigment represents more hydrophobic residues (positive values), while white and cyan colors represent neutral and hydrophilic (negative values), respectively.

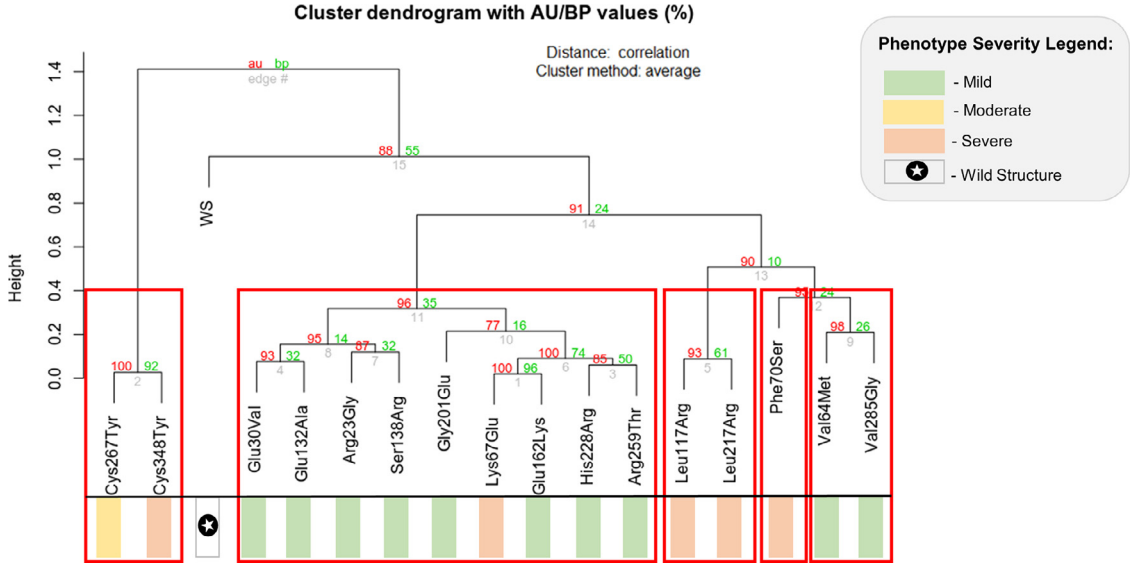

Fig. 4. Hierarchical clustering analysis (HCA) of combining physicochemical properties values obtained from mutated A1 domain structures against the wild type. The conducted analysis uses correlation distance and average clustering method, calculating approximately unbiased (AU) values based on 10,000 bootstraps (BP) replications. We consider an AU larger than 95% percent ($P < 0{,}0001$) to form clusters. The clustering considered variants and the wild-type structure combining values of all analyzed physicochemical features and substitution matrices. The variant phenotypes colors are in the legend: green, yellow, and red for the mild, moderate, and severe phenotypes, respectively. The wild-type structure (WS) legend contains an inside-circle star.

shows only one instance where a mild and a severe variant grouped together (p.Glu162Lys and p.Lys67Glu), the HCA considering only substitutions indexes shows three cases: p.Lys67Glu (severe)/p.Glu162Lys(mild); p.Phe70Ser(severe)/ p.Val285Gly(mild); and p.Val64met(mild)/p.Gly201Glu(severe).

3.3. Principal component analysis

For variants in the A1 domain, the first principal component (PC) represents 34.63% of the total variance, while the second one contributes 22. 52% (Fig. 5**A** and Table 2). Together, the first two PCs explain 57% of the data variance, indicating their essential role. The five first dimensions pertaining to variants in domain A1 explain almost 91% of the variance, as presented in Table 2. Table S2 presents the coordinate and contributions for elements in domains A2, A3, C1, and C2, with the five first dimensions results explaining respectively 90.07, 91.02, 100, and 96.92% of variance. Fig. 5**B** shows the PCA results for the different variables, and for the individual A1 variants, considering the two most representative dimen-

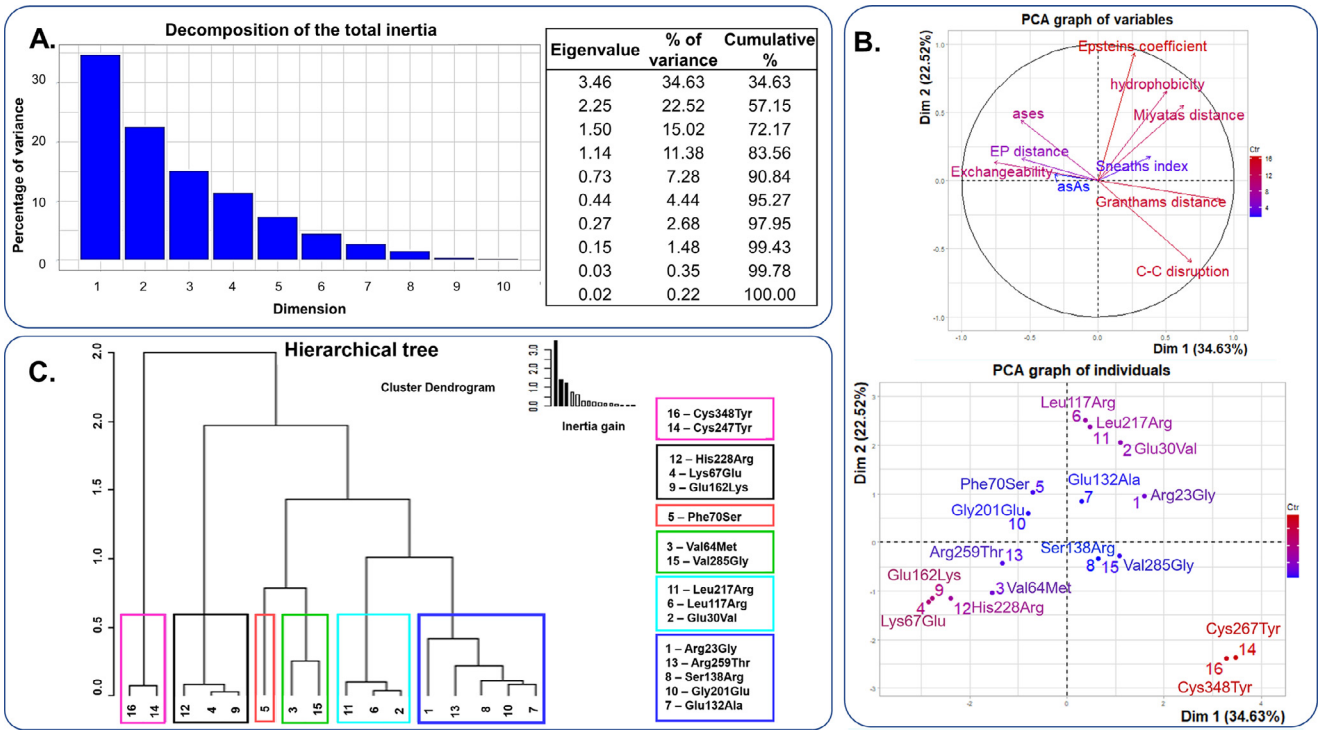

Fig. 5. Principal component analysis (PCA) of physicochemical properties in the A1 domain. In A - Graphic showing decomposition of total inertia and percentage of variance in the ten dimensions. The table shows the eigenvalues, variance percentage, and cumulative percentage of variance. **B**- Two-dimensional graphics for variables and individuals. Colors of eigenvectors and variables are in conformity to scale (red for more contributive properties, and blue for less). "Ctr" means contribution. Individuals' distribution according to the features. **C-** Hierarchical tree grouping the substitutions according to the principal components.

Table 2
Variables' contributions and coordinates considering their first five dimensions for A1 domain mutations.

	Contributions					Coordinates				
Variables	*Dim*[†]*.1*	*Dim.2*	*Dim.3*	*Dim.4*	*Dim.5*	*Dim.1*	*Dim.2*	*Dim.3*	*Dim.4*	*Dim.5*
AreaSAS	2.91	0.09	0.00	69.08	12.99	−0.32	0.05	−0.00	0.89	0.31
AreaSES	9.25	8.66	13.79	2.79	22.75	−0.57	0.44	0.46	−0.18	0.41
C-C disruption	13.60	15.85	0.42	0.26	9.10	0.69	−0.60	0.08	0.05	0.26
EP distance	9.00	1.16	32.22	2.39	0.56	−0.56	0.16	0.70	−0.16	−0.06
Epsteins coefficient	2.06	38.76	0.89	0.49	1.94	0.27	0.93	−0.12	0.07	0.12
Exchangeability	16.75	0.79	5.70	0.08	19.63	−0.76	0.13	−0.29	−0.03	−0.38
Granthans distance	23.41	0.88	6.38	1.71	0.00	0.90	−0.14	0.31	−0.14	0.01
Hydrophobicity	7.31	19.07	5.51	9.13	6.66	0.50	0.66	−0.29	−0.32	0.22
Miyatas distance	11.40	13.41	0.04	10.17	15.62	0.63	0.55	−0.02	0.34	−0.34
Sneaths index	4.28	0.79	5.70	0.08	19.63	0.38	0.17	0.73	0.21	−0.28

[†] The abreviation *Dim* is used as mean for *Dimension*.The dimensions contributions and coordinates values were obtained using FactorMiner package in R studio cross-platform.

sions. The same analysis was performed for variants in domains A2, A3, C1 and C2, and the results are presented in Fig. S15. Hierarchical clustering for Principal components (HCPC) was performed with Euclidean metrics (*p*-value of 0.05) for A1 variants, revealing six clusters based on the first five dimensions. HCPC shows a distinct pattern from HCA: three clusters group mild phenotype variants together with severe ones: p.Glu30Val(mild) with p.Leu117Arg and p.Leu217Arg; p.Gly201Glu (severe) with four mild variants; p.Lys67Glu (severe) with p.His228Arg and p.Glu162Lys (mild).

4. Discussion

4.1. FVIII electrostatic potential property is impacted in more than 60% of hemophilia A missense variants

Electrostatic potential (EP) consists of a fundamental property used to infer disruption or gain of affinity in protein-protein interactions [25,26], in the particular case of FVIII, its interaction with coagulation cascade partners, such as the Factor IX. Most of the current predictive tools do not consider EP at all, or if they do, they are limited to residue changes and not applied to the calculated surface (from domain or protein) as a whole [27]. A change in EP can produce a more profound impact in the protein, which might not be observed only by local comparisons between wild-type and substituted amino acids [28]. EP disruptions caused by missense variants in FVIII set their respective models apart from the wild-type phenotype in the epigrams, but influence on disease phenotype varied according to domain affected and the nature of substitutions (Figs. 2**, S1–S4**). Also, visual inspection of these models demonstrated that the charge changes are distributed across FVIII domains and throughout its surface, which is hard to detect in a two-dimensional visualization scenario showing only one face of the protein (as represented in Fig. 2). The percentage of substitutions that were distinct from WS (about 69% of the total, and 65.7% among severe phenotype variants only) is consistent with the cru-

cial role of EP, but also indicates the participation of other features in phenotype determination.

4.2. Amino acid substitutions affecting properties critical for protein function: all pathogenic models show changes in at least one physicochemical attribute

A residue's affinity for water consists of a molecular characteristic capable of modifying solvent exposure, leading to structural changes in the protein, and affecting its function and dynamics [29,30]. The high number of substitutions causing hydrophobicity changes highlights the role of this feature in determining the effect of missense variants, despite the fact that the hydrophobicity dimension is not linearly related to phenotype. Even substitutions with no effect on EP were shown to have effects on hydrophobicity, which was related to the properties of the substituted amino acids regardless of position.

Disruption od disulfide bonds seems to be a good predictor of functional impact because it can directly affect formation and stability of protein domains. Most likely, the loss of cysteine residue determines the outcome, rather than the nature of the replacing amino acid [31,32]. Unsurprisingly, none of the pathogenic variants which had disulfide disruption were of mild phenotype (Table 1).

Modifications in the solvent-accessible and solvent-excluded surface areas (areaSAS and areaSES, respectively) occur because of a high interference with the molecular surface, turning the residues more enclosed (or exposed) [33,34]. These features are related to the surface, not having an amino acid-specific predefined value, so the position in the protein chain and neighbor residues also contribute to the outcome. This is corroborated by our observations, for example, for variants p.Leu117Arg and p.Leu217Arg attributes values. The sum of aSAS and aSES of all residues in each model are of 419.269.401.417.594 (aSAS) and 420.174.296.508.431 (aSES) for Arg117, and 419.223.427.798.033 (aSAS) and 420.038.242.070.079 (aSES) for Arg217. Even though the same substitution is seen in both cases, it is possible to observe the difference in area values, as is the case for all other instances of identical substitution in different positions (Table S1).

4.3. The use of amino acid replacement indexes improves predictions, in a combined approach

The replacement distance indexes consist of mathematical calculations to estimate values for the difference between amino acids considering their properties, with each index examining a different property. In this context, it does not seem correct to consider any of them separately, which could increase the estimated weight bias of the physicochemical property effect at the expense of others. The factors considered by each index are not identical [35], so the combined use of them seems to be more effective. Importantly, when observed and compared to patient phenotypes (without considering other properties), these index values do not explain the severity of pathogenic variants described, nor singly or in combination (Fig. S10 shows an HCA for A1 domain using only the substitutions matrices data). Nevertheless, including the substitution matrices improved the refinement of the HCA approach (combining all properties).

4.4. Using multiple physicochemical features in a hierarchical clustering analysis substantially reduced the number of variants grouped with the wild-type structure

The HCA based on the combination of the different physicochemical aspects analyzed here yields more consistent clusters in comparison with those based on EP or substitutions indexes isolatedly. In the HCA results for the A1 domain (Fig. 4), besides the accuracy in distinguishing WS from the mutated structures, we can also observe a greater homogeneity of clusters' components regarding phenotypes. There is only a single instance when a model corresponding to a severe phenotype (p.Lys67Glu) was placed within a cluster of mild variants. Interestingly, this case involves residue substitution from a positively-charged Lys-to a negatively-charged Glu. The observed improvement in HCA resolution is also seen for the other domains (Figs. S11–S14), but it is not possible to confirm a straight effect associated with severity. Altogether, the analyses highlight the importance of the considered physicochemical features as phenotypic determining aspects in HA (and probably other Mendelian disorders).

4.5. A multivariate principal component analysis helps elucidating the role of different features in specific domains

Principal component analysis (PCA) is valuable in clarifying the directions where variance is more pronounced [36]. This linear transformation fits information in a coordinate system, where related variants contribute with similar property values as a combined component [37]. In a two-dimensional graphic (Fig. 5**B**), the eigenvalues depiction reproduces the eigenvectors force. The variant distribution in this coordinate system tends to group in conformity with characterized elements. Considering the PCA graph of individuals A1 domain variants, the eight most contributive variants (Fig. 5**B**) are distributed according to variable effects. The substitutions p.Cys267Tyr and p.Cys348Tyr (pink cluster) are highly influenced by the eigenvector "C-C disruption"; while another group (black cluster: p.Lys67Glu, p.Glu162Lys, and p.His228Arg) position is related to distinct features (Granthams, Exchangeability, areaSES); the remaining three variants (light blue cluster: p.Leu117Arg, p.Leu217Arg, and p.Glu30Val) are characterized by high values for Epstein coefficient, hydrophobicity, and Miyata. This data corroborates the qualitative and quantitative observed hydrophobicity changes (Fig. 3) observed for p.Leu117Arg, p.Leu217Arg, and p.Glu30Val, which have a more prominent change than all other evaluated substitutions in the same domain, which were more centrally located in the PCA graphs, indicating to be less influenced by these analyzed.

The HCPC analysis (Fig. 5**C**) shows a poorer performance compared with the HCA (Fig. 4). It demonstrates that the use of only the five most contributive features extracted from PCA did not improve phenotype grouping. In turn, it pointed out the importance of a broad approach including the whole set of available variables, highlighting the concept that every feature has a role in clusterization. Analyzing the remaining FVIII domains, it is possible to notice distinct forces and contribution of components according to the features determined by variants, which occur due to the particular nature of a given substitution and have a domain specific effect.

The five most contributive properties varied among the different domains: for example, the A3 domain substitutions showed different variables coordinate distribution in comparison with C1 (Fig. S15B and **S15C**). On the A3 domain, EP has a major impact in the first two PCs, which helps to explain the more similar results found between EP clustering (Fig. S2) and HCA (Fig. S12) for A3. The distance scale depicted in A3 electrostatic distribution clustering (Fig. S2) is more prominent (0–0.8) when compared with other domains (0–0.25 for A1, A2, C1, and 0–0.5 for C2), highlighting a more divergent EP profile. Considering the A3 domain functionality, the consequences of EP alterations could be determinant due to the domain's important role in activation (proximity and participation on B domain cleavage) and stabilization (linkage with Von Willebrand Factor) [38]. EP changes possibly affect the affinity required among the involved interacting residues on the protein surface. No A3 variant showed disrupted disulfide bonds, and

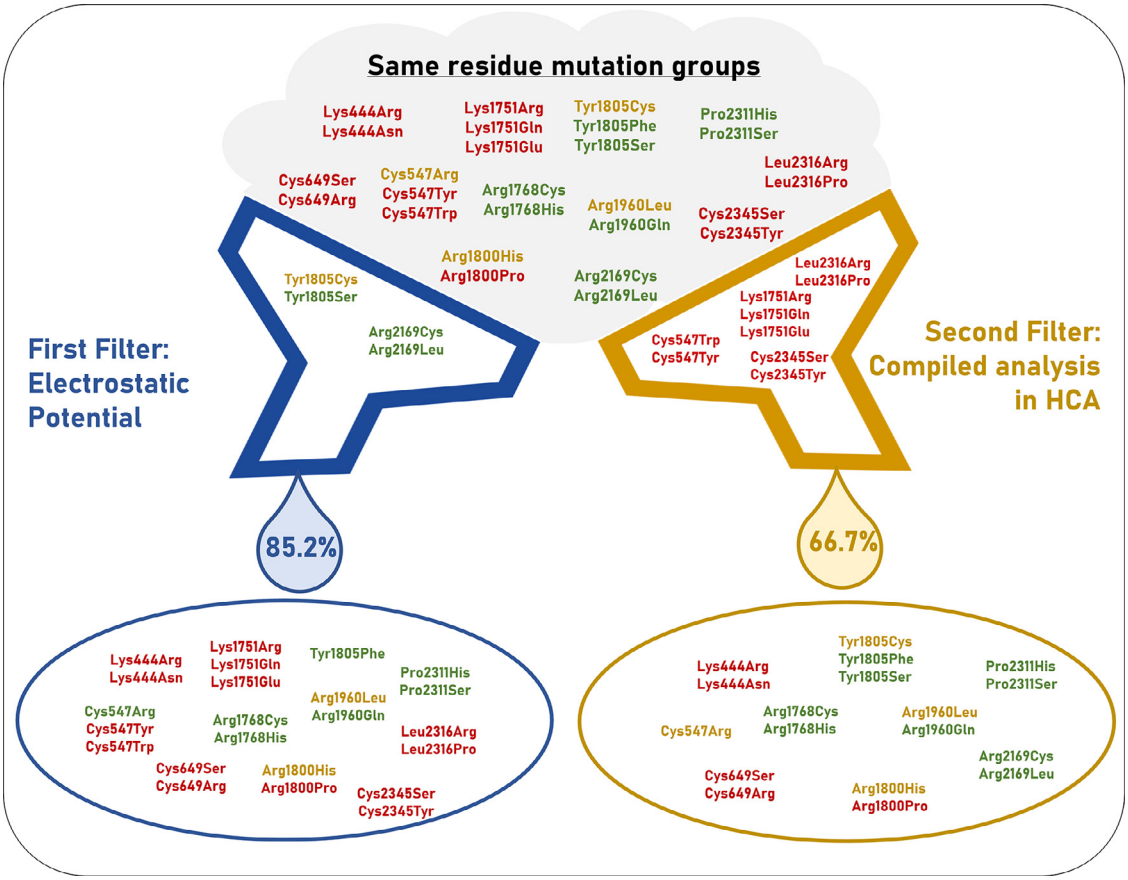

Fig. 6. Separate variants in given wild-type residues in electrostatic potential terms and combined properties in HCA. On the top cloud, a total of 27 variants occur in given wild-type residues. The variant' colors represent their phenotype: green, yellow, and red for mild, moderate, and severe, respectively. The variants passed through one filter evaluating EP (blue), and the other involved HCA analysis combining physicochemical properties (yellow). The variants on a given position that group together remain stuck in the barriers, and those apart transposed the obstacle. About 85% of variants passed through the EP funnel (blue), and about 67% transposed the combined properties funnel (yellow). Substitutions that cross the barriers are in the spheres, according to their color groups.

thus this feature does not appear among the contributing factors. On the other hand, hydrophobicity and disulfide-bond disruptions have a higher impact on the C1 domain, which includes the cysteine changes in residues 2040 and 2188 involved in domain stability. As expected, the results of a PCA including all 71 variants for the five active FVIII domains (A1, A2, A3, C1, and C2) (Fig. S16) also differ from the analysis of the individual domains.

4.6. Each variant effect is determined by changes in properties and by the substitution position

The variants with substitutions in the same sequence position presented alternative clustering depending on whether EP alone or the combined properties were considered for HCA. Fig. 6 illustrates these variants that are grouped together(retained) after each clustering analysis. While the results are discrepant, again the combined use of physicochemical properties including EP improved the clustering of variants according to their phenotype. Cysteine bond disruptions are considered as a determining aspect in FVIII structure stability explaining the tendency to retain together variants impacted by changes in this feature. The variant p.Cys547Arg (moderate phenotype) is separated from p.Cys547Tyr and p.Cys547Trp (both severe phenotypes) in the HCA (Figs. 6 and S15), strengthening the PCA results that pointed to more properties influencing the p.Cys547Arg substitution (as Miyata's distance, Sneath's index, Granthamś distance). Finally, these findings emphasize that the effect of different variants on phenotype depend on the properties of substituted amino acids, as well as their position within the protein, and consequent changes in important features of the wild-type structure.

4.7. Analysis of non-pathogenic p. Glu132Asp polymorphism supports the role of FVIII physicochemical properties in HA determination

Additionally, we conducted a search for F8 polymorphisms and found one non-conservative substitution considered non-pathogenic in the A1 domain, p.Glu132Asp [39] which appears in a non-hemophilia screen [40]. Fig. 7 presents the result of hierarchical clustering including p.Glu132Asp. The non-pathogenic polymorphism appeared in the same branch with wild-type structure (with no other pathogenic variant), revealing that despite its physicochemical alterations (which include mild modifications in EP, aSAS, aSES, and substitutions matrices), the variant is more closely related to WS than to pathogenic variants. The inclusion of a large dataset of FVIII polymorphisms could be applied in a machine learning method in a future approach. In this sense, AI methods considering the points defined in the present work will allow the generation of predictors and classifiers which could be used to accurately infer clinical aspects in new emergent variants. Moreover, the applied approach developed here also contributes to understanding mechanisms that lead to changes in protein function, and which protein sites are important for molecular interactions.

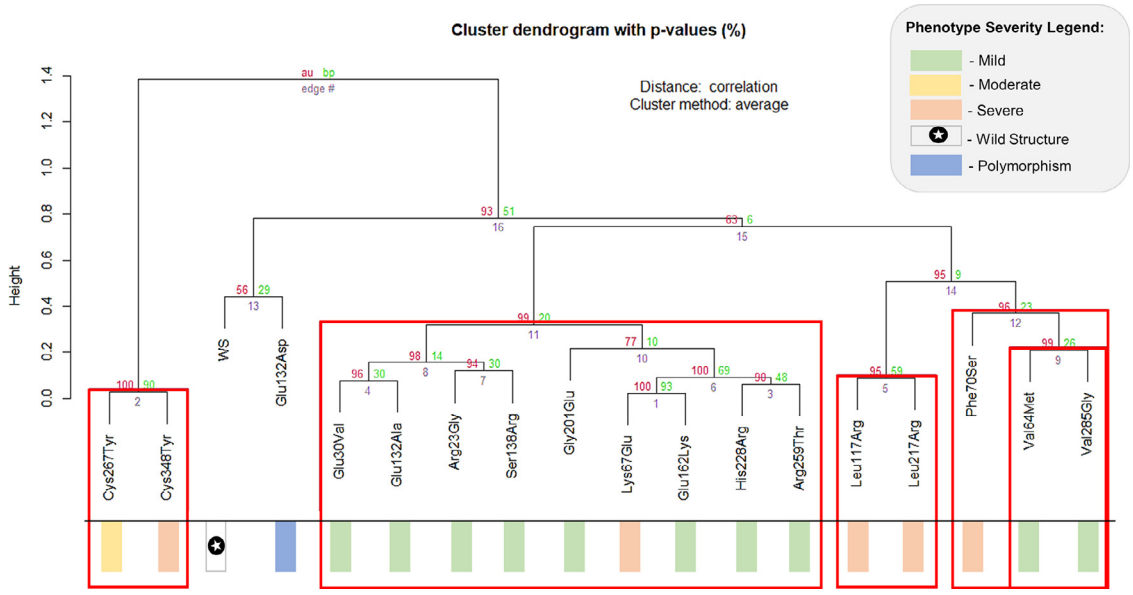

Fig. 7. Inclusion of non-pathogenic polymorphic variant p.Glu132Ala in the hierarchical clustering of A1 domain. The hierarchical clustering combining the previously mentioned features groups the non-pathogenic variant p.Glu132Asp more closely to wild-type, revealing that the combination of physicochemical properties distinguishes it from the pathogenic variants.

5. Conclusions

Previous studies already investigated the effects of missense variants on disease phenotype using *in silico* approaches [41–44]. The present work brings novel insights through a structural bioinformatics pipeline in hemophilia A variants, with the potential to be applied to other Mendelian disorders. The physicochemical changes generated by missense variants influence disease determination in a complex way [45]. None of the aspects evaluated here could individually explain the phenotype, however, the combined use of this set of characteristics in an HCA proved to be an effective method to discriminate variants according to amino acid properties, since substitutions have multiple consequences at the protein level. Every feature proved to be important to explain the condition.

The PCA approach revealed which variables had the most predominant impact, bringing up an essential issue about predictions: domains specificities. As observed in our data, a mutational landscape sometimes leads to different phenotypes according to domain functionality. So, to increase accuracy, prediction methods must contemplate domains and their influence on protein function. The mutational spectrum tolerated by each protein region is unique, and consequently, so is the effect in determining disorder severity.

Inferring the effects of different variants on severity seems a bigger challenge, since there are features with continuous distribution (i.e. EP, aSES, aSAS), of which estimations consider the presented surface as a whole (domain or protein). Other aspects are calculated based on penalty score matrices that indicate solely the effect of residue change. Combining the impact caused by each physicochemical alteration is crucial to develop a more accurate tool to predict phenotypic outcomes (molecular changes that affect protein function and secretion into plasma).

The pre-established values of some features (disulfide-disruption, hydrophobicity, index matrices) based on the amino acid properties result in a higher similarity between variants where identical amino acids substitutions occur in distinct protein sites (as can be noticed in Fig. 3 for p.Cys267Tyr and p.Cys348Tyr, where the same hydrophobicity values are assigned - from 2.5 to −1.3 for both sites). EP, aSES and aSAS are related to the surface area modifications turning the analysis more complex, at the same time allowing a better understanding of position-related outcomes, because they involve flanking residues besides the substituted amino acids. Changes at the same wild-type residue, and more importantly, conserve all other amino acids in the chain, so divergences are controlled and relationships can be established. Alternatively, creating models for variants with all possible amino acid substitutions in a given site enables the evaluation of which models are more closely related to the wild type. This option yields interpretations, but is time-consuming and massive in computational demands. This approach becomes more feasible after the identification of pivotal elements for each considered system, as observed in the present study for FVIII protein. In this sense, female carriers of *F8* gene mutations could benefit from the presented approach, which tends to reveal how amino acids modifications interfere in protein function and disease severity. It could be implemented in genetic counseling to guide decisions. Also, considering the current development in this field, variant information is very useful for biotechnology approaches, experimental design, as well as synthetic biology and genetic editing.

Declaration of Competing Interest

None

CRediT authorship contribution statement

Mariana R. Meireles: Conceptualization, Methodology, Software, Formal analysis, Investigation, Writing – original draft, Writing – review & editing. **Lara H. Stelmach:** Formal analysis, Investigation. **Eliane Bandinelli:** Conceptualization, Writing – review & editing, Supervision. **Gustavo F. Vieira:** Conceptualization, Writing – review & editing, Supervision.

Acknowledgments

Financing for this work was provided by Conselho Nacional de Desenvolvimento Científico e Tecnológico and Fundação de Amparo

à Pesquisa do Estado do Rio Grande do Sul (Apoio a Núcleos de Excelência Program).

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.cmpb.2022.106768.

References

[1] L.W. Hoyer, A hemophilia, N. Engl. J. Med. 330 (1994) 38–47, doi:10.1056/NEJM199401063300108.

[2] J. Gitschier, et al., Characterization of the human factor VIII gene, Nature 312 (1984) 326–330.

[3] D.L. Eaton, G.A. Vehar, Factor VIII structure and proteolytic processing, Prog. Hemost. Thromb. 8 (1986) 47–70.

[4] G.A. Vehar, et al., Structure of human factor VIII, Nature 312 (1984) 337–342.

[5] P.J. Lenting, J.A. van Mourik, K. Mertens, The life cycle of coagulation factor VIII in view of its structure and function, Blood 92 (1998) 3983–3996.

[6] J.C. Ngo, M. Huang, D.A. Roth, B.C. Furie, B. Furie, Crystal structure of human factor VIII: implications for the formation of the factor IXa-factor VIIIa complex, Structure 16 (2008) 597–606, doi:10.1016/j.str.2008.03.001.

[7] G.C. White, et al., Definitions in hemophilia. Recommendation of the scientific subcommittee on factor VIII and factor IX of the scientific and standardization committee of the International society on thrombosis and Haemostasis, Thromb. Haemost. 85 (2001) 560.

[8] D. Venkateswarlu, Structural insights into the interaction of blood coagulation co-factor VIIIa with factor IXa: a computational protein-protein docking and molecular dynamics refinement study, Biochem. Biophys. Res. Commun. 452 (2014) 408–414, doi:10.1016/j.bbrc.2014.08.078.

[9] P.H. Bolton-Maggs, K.J. Pasi, Haemophilias A and B, Lancet 361 (2003) 1801–1809, doi:10.1016/S0140-6736(03)13405-8.

[10] L.A. Kelley, S. Mezulis, C.M. Yates, M.N. Wass, M.J. Sternberg, The phyre2 web portal for protein modeling, prediction and analysis, Nat. Protoc. 10 (2015) 845–858, doi:10.1038/nprot.2015.053.

[11] A.H.A. Maghrabi, L.J. McGuffin, ModFOLD6: an accurate web server for the global and local quality estimation of 3D protein models, Nucleic Acids Res. 45 (2017) W416–W421, doi:10.1093/nar/gkx332.

[12] L.J. McGuffin, M.T. Buenavista, D.B. Roche, The ModFOLD4 server for the quality assessment of 3D protein models, Nucleic Acids Res. 41 (2013) W368–W372, doi:10.1093/nar/gkt294.

[13] R.A. Laskowski, et al., PDBsum: a web-based database of summaries and analyses of all PDB structures, Trends Biochem. Sci. 22 (1997) 488–490.

[14] S. Sarkar, et al., DelPhi web server: a comprehensive online suite for electrostatic calculations of biological macromolecules and their complexes, Commun. Comput. Phys. 13 (2013) 269–284.

[15] N. Smith, et al., DelPhi web server v2: incorporating atomic-style geometrical figures into the computational protocol, Bioinformatics 28 (2012) 1655–1657, doi:10.1093/bioinformatics/bts200.

[16] E.F. Pettersen, et al., UCSF chimera–a visualization system for exploratory research and analysis, J. Comput. Chem. 25 (2004) 1605–1612, doi:10.1002/jcc.20084.

[17] S. Richter, A. Wenzel, M. Stein, R.R. Gabdoulline, R.C. Wade, webPIPSA: a web server for the comparison of protein interaction properties, Nucleic Acids Res. 36 (2008) W276–W280, doi:10.1093/nar/gkn181.

[18] R. Grantham, Amino acid difference formula to help explain protein evolution, Science 185 (1974) 862–864, doi:10.1126/science.185.4154.862.

[19] P.H. Sneath, Relations between chemical structure and biological activity in peptides, J. Theor. Biol. 12 (1966) 157–195, doi:10.1016/0022-5193(66)90112-3.

[20] C.J. Epstein, Non-randomness of amino-acid changes in the evolution of homologous proteins, Nature 215 (1967) 355–359, doi:10.1038/215355a0.

[21] T. Miyata, S. Miyazawa, T. Yasunaga, Two types of amino acid substitutions in protein evolution, J. Mol. Evol. 12 (1979) 219–236, doi:10.1007/bf01732340.

[22] L.Y. Yampolsky, A. Stoltzfus, The exchangeability of amino acids in proteins, Genetics 170 (2005) 1459–1472, doi:10.1534/genetics.104.039107.

[23] R. Suzuki, H. Shimodaira, Pvclust: an R package for assessing the uncertainty in hierarchical clustering, Bioinformatics 22 (2006) 1540–1542, doi:10.1093/bioinformatics/btl117.

[24] S. Lê, J. Josse & F. Husson FactoMineR: an R package for multivariate analysis. 2008**25**, 18, doi:10.18637/jss.v025.i01 (2008).

[25] S.M. Ivanov, A. Cawley, R.G. Huber, P.J. Bond, J. Warwicker, Protein-protein interactions in paralogues: electrostatics modulates specificity on a conserved steric scaffold, PLoS One 12 (2017) e0185928, doi:10.1371/journal.pone.0185928.

[26] H.X. Zhou, X. Pang, Electrostatic interactions in protein structure, folding, binding, and condensation, Chem. Rev. 118 (2018) 1691–1741, doi:10.1021/acs.chemrev.7b00305.

[27] H. Tang, P.D. Thomas, Tools for predicting the functional impact of nonsynonymous genetic variation, Genetics 203 (2016) 635–647, doi:10.1534/genetics.116.190033.

[28] X. Yang, et al., On the role of sidechain size and charge in the aggregation of A β 42 with familial mutations, Proc. Natl. Acad. Sci. 115 (2018) E5849–E5858, doi:10.1073/pnas.1803539115.

[29] J. Dygut, et al., Structural interface forms and their involvement in stabilization of multidomain proteins or protein complexes, Int. J. Mol. Sci. 17 (2016), doi:10.3390/ijms17101741.

[30] M. Banach, L. Konieczny, I. Roterman, Why do antifreeze proteins require a solenoid? Biochimie 144 (2018) 74–84, doi:10.1016/j.biochi.2017.10.011.

[31] T. Liu, et al., Enhancing protein stability with extended disulfide bonds, Proc. Natl. Acad. Sci. USA 113 (2016) 5910–5915, doi:10.1073/pnas.1605363113.

[32] W.J. Wedemeyer, E. Welker, M. Narayan, H.A. Scheraga, Disulfide bonds and protein folding, Biochemistry 39 (2000) 7032, doi:10.1021/bi005111p.

[33] J.A. Marsh, Buried and accessible surface area control intrinsic protein flexibility, J. Mol. Biol. 425 (2013) 3250–3263, doi:10.1016/j.jmb.2013.06.019.

[34] M.A. Moret, G.F. Zebende, Amino acid hydrophobicity and accessible surface area, Phys. Rev. E Stat. Nonlin. Soft. Matter. Phys. 75 (2007) 011920, doi:10.1103/PhysRevE.75.011920.

[35] T. Dagan, Y. Talmor, D. Graur, Ratios of radical to conservative amino acid replacement are affected by mutational and compositional factors and may not be indicative of positive Darwinian selection, Mol. Biol. Evol. 19 (2002) 1022–1025, doi:10.1093/oxfordjournals.molbev.a004161.

[36] I.T. Jolliffe, J. Cadima, Principal component analysis: a review and recent developments, Philos. Trans. A Math. Phys. Eng. Sci. 374 (2016) 20150202, doi:10.1098/rsta.2015.0202.

[37] M. Ringnér, What is principal component analysis? Nat. Biotechnol. 26 (2008) 303–304, doi:10.1038/nbt0308-303.

[38] E. Bloem, H. Meems, M. van den Biggelaar, K. Mertens, A.B. Meijer, A3 domain region 1803-1818 contributes to the stability of activated factor VIII and includes a binding site for activated factor IX, J. Biol. Chem. 288 (2013) 26105–26111, doi:10.1074/jbc.M113.500884.

[39] K. Freson, et al., Fluorescent chemical cleavage of mismatches for efficient screening of the factor VIII gene, Hum. Mutat. 11 (1998) 470–479, doi:10.1002/(SICI)1098-1004(1998)11:6(470::AID-HUMU8)3.0.CO;2-A.

[40] P.S. Tarpey, et al., A systematic, large-scale resequencing screen of X-chromosome coding exons in mental retardation, Nat. Genet. 41 (2009) 535–543, doi:10.1038/ng.367.

[41] K. Réblová, et al., Hyperphenylalaninemia in the Czech Republic: genotype-phenotype correlations and *in silico* analysis of novel missense mutations, Clin. Chim. Acta 419 (2013) 1–10, doi:10.1016/j.cca.2013.01.006.

[42] T.E. Heineman, D.G. Evans, F. Campagne, S.H.In Selesnick, *Silico* analysis of NF2 gene missense mutations in neurofibromatosis type 2: from genotype to phenotype, Otol. Neurotol. 36 (2015) 908–914, doi:10.1097/MAO.0000000000000639.

[43] M.R. Meireles, M.A.S. Bragatte, E. Bandinelli, F.M. Salzano, G.F. Vieira, A new *in silico* approach to investigate molecular aspects of factor IX missense causative mutations and their impact on the hemophilia B severity, Hum. Mutat. 40 (2019) 706–715, doi:10.1002/humu.23733.

[44] L. Meléndez-Aranda, A.R. Jaloma-Cruz, N. Pastor, M.M.J. Romero-Prado, *In silico* analysis of missense mutations in exons 1-5 of the F9 gene that cause hemophilia B, BMC Bioinf. 20 (2019) 363, doi:10.1186/s12859-019-2919-x.

[45] C.M. Yates, M.J. Sternberg, The effects of non-synonymous single nucleotide polymorphisms (nsSNPs) on protein-protein interactions, J. Mol. Biol. 425 (2013) 3949–3963, doi:10.1016/j.jmb.2013.07.012.