Computational and Structural Biotechnology Journal 21 (2023) 5249–5258

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

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

Research article

A multi-layered computational structural genomics approach enhances domain-specific interpretation of Kleefstra syndrome variants in EHMT1

Young-In Chi [a,b], Salomão D. Jorge [a], Davin R. Jensen [a,c], Brian C. Smith [a,c], Brian F. Volkman [a,c], Angela J. Mathison [a,b], Gwen Lomberk [a,b,d], Michael T. Zimmermann [a,c,e,*], Raul Urrutia [a,b,c,*]

[a] Linda T. and John A. Mellowes Center for Genomic Sciences and Precision Medicine, Medical College of Wisconsin, Milwaukee, WI, USA
[b] Division of Research, Department of Surgery, Medical College of Wisconsin, Milwaukee, WI, USA
[c] Department of Biochemistry, Medical College of Wisconsin, Milwaukee, WI, USA
[d] Department of Pharmacology and Toxicology, Medical College of Wisconsin, Milwaukee, WI, USA
[e] Clinical and Translational Sciences Institute, Medical College of Wisconsin, Milwaukee, WI, USA

ARTICLE INFO

Keywords:
EHMT1
GLP
Epigenetic regulator
Histone methyltransferase
Gene variation
Kleefstra syndrome
SET domain
Mutational impact analysis
Molecular dynamics

ABSTRACT

This study investigates the functional significance of assorted variants of uncertain significance (VUS) in euchromatic histone lysine methyltransferase 1 (EHMT1), which is critical for early development and normal physiology. EHMT1 mutations cause Kleefstra syndrome and are linked to various human cancers. However, accurate functional interpretations of these variants are yet to be made, limiting diagnoses and future research. To overcome this, we integrate conventional tools for variant calling with computational biophysics and biochemistry to conduct multi-layered mechanistic analyses of the SET catalytic domain of EHMT1, which is critical for this protein function. We use molecular mechanics and molecular dynamics (MD)-based metrics to analyze the SET domain structure and functional motions resulting from 97 Kleefstra syndrome missense variants within the domain. Our approach allows us to classify the variants in a mechanistic manner into SV (Structural Variant), DV (Dynamic Variant), SDV (Structural and Dynamic Variant), and VUS (Variant of Uncertain Significance). Our findings reveal that the damaging variants are mostly mapped around the active site, substrate binding site, and pre-SET regions. Overall, we report an improvement for this method over conventional tools for variant interpretation and simultaneously provide a molecular mechanism for variant dysfunction.

1. Introduction

Over the last three decades, extensive work has recognized that histone modifications are central to epigenetic regulation. Epigenetic dysregulation caused by mutations in components of histone-modifying enzymes leads to various human diseases known as chromatinopathies [1]. EHMT1, also called G9a-like protein (GLP), catalyzes mono- and di-methylation of Lys9 of histone H3 (H3K9me1 and H3K9me2) for gene silencing [2]. EHMT1 alterations are associated with Kleefstra syndrome (OMIM 610253), a neurodevelopmental disorder, and different tumor types, including uterine, adrenocortical and skin melanoma, and stomach adenocarcinoma [3]. To benefit patients with cancer and suspected chromatinopathies, improved methods are necessary to interpret EHMT1 genetic alterations.

Belonging to the SET domain-containing methyltransferase family, the human EHMT1 protein consists of 1298 amino acids and is characterized by distinctive domains, including the transactivation domain at the *N*-terminus, the cysteine-rich domain, the ankyrin repeat domain

Abbreviations: **COSMIC**, Catalogue of Somatic Mutations in Cancer; **dbSNP,**, Single nucleotide polymorphism database; **DV,**, Dynamic variant; **EHMT1,**, Euchromatic histone lysine methyltransferase 1; **GLP,**, G9a-like protein; **gnomAD,**, genome aggregation database; **KDM6A,**, Lysine-specific demethylase 6A; **MD,**, Molecular dynamics; **OMIM,**, Online Mendelian inheritance in man; **PDB,**, Protein data bank; **Rg,**, Radius of gyration; **RMSD,**, Root mean square deviation; **RMSF,**, Root mean square fluctuation; **SAH,**, S-adenosylhomocysteine; **SAM,**, S-adenosylmethionine; **SASA,**, Solvent-accessible surface area; **SDV,**, Structural and dynamic variants; **SET,**, Su(var)3–9, Enhancer-of-zeste and Trithorax; **SNP,**, Single nucleotide polymorphism; **SV,**, Structural variant; **VUS,**, Variant of uncertain (unknown) significance.

* Corresponding authors at: Linda T. and John A. Mellowes Center for Genomic Sciences and Precision Medicine, Medical College of Wisconsin, Milwaukee, WI, USA.

E-mail addresses: mtzimmermann@mcw.edu (M.T. Zimmermann), rurrutia@mcw.edu (R. Urrutia).

https://doi.org/10.1016/j.csbj.2023.10.022
Received 8 June 2023; Received in revised form 6 October 2023; Accepted 12 October 2023
Available online 13 October 2023

with scaffolding function, and the enzymatic SET domain at the *C*-terminus. Responsible for the writer function of EHMT1, the SET domain catalyzes mono- and di-methylation of the H3K9 residue. Reader function is conferred by the ankyrin repeat domain, which recognizes the same histone modification (H3K9me1/2) and mediates higher-order complex formations for gene and epigenome regulations [4]. Thus, although the full-length protein cohesively exerts its function, the impact of mutations on domain-specific molecular mechanisms should be considered, as each domain represents an independent folding unit and possesses a discrete function. While numerous Kleefstra syndrome germline variants have been identified in patients [5,6], their pathogenicity and dysfunctional molecular defects are poorly understood. Current variant interpretations rely heavily on 2D sequence-based information and limited structural and functional data [7]. Thus, computational approaches and multiplexed experimental data are urgently needed to improve prediction power and delineate potential dysfunctional mechanisms [8].

This study reports a domain-wide analysis of SET catalytic domain variants associated with the Kleefstra syndrome congenital disease to fill this knowledge gap. The SET catalytic domain variants have unique molecular features that are shared by its homolog EHMT2 and other members of the SET domain-containing methyltransferases. We studied 97 missense variants (on 82 residues) within the EHMT1 SET domain using the available crystal structure (PDB access code 3HNA) to predict their impacts on the structural and dynamic properties of the protein. We applied selective analytical tools from each protein layer representing universal or protein-specific and global or local considerations, including folding/stability energy, structure perturbation, binding energy calculations, local geometry analyses, and all-atom MD simulations. These multi-tiered mechanistic-based analyses complement existing prediction tools and further enhance the mutational impact assessments highly relevant to protein structure and function. Thus, this study represents a novel approach to understanding the functional effects of these alterations by providing a broader characterization of genomic variants with dynamic modeling specific to a rare disease-associated SET domain.

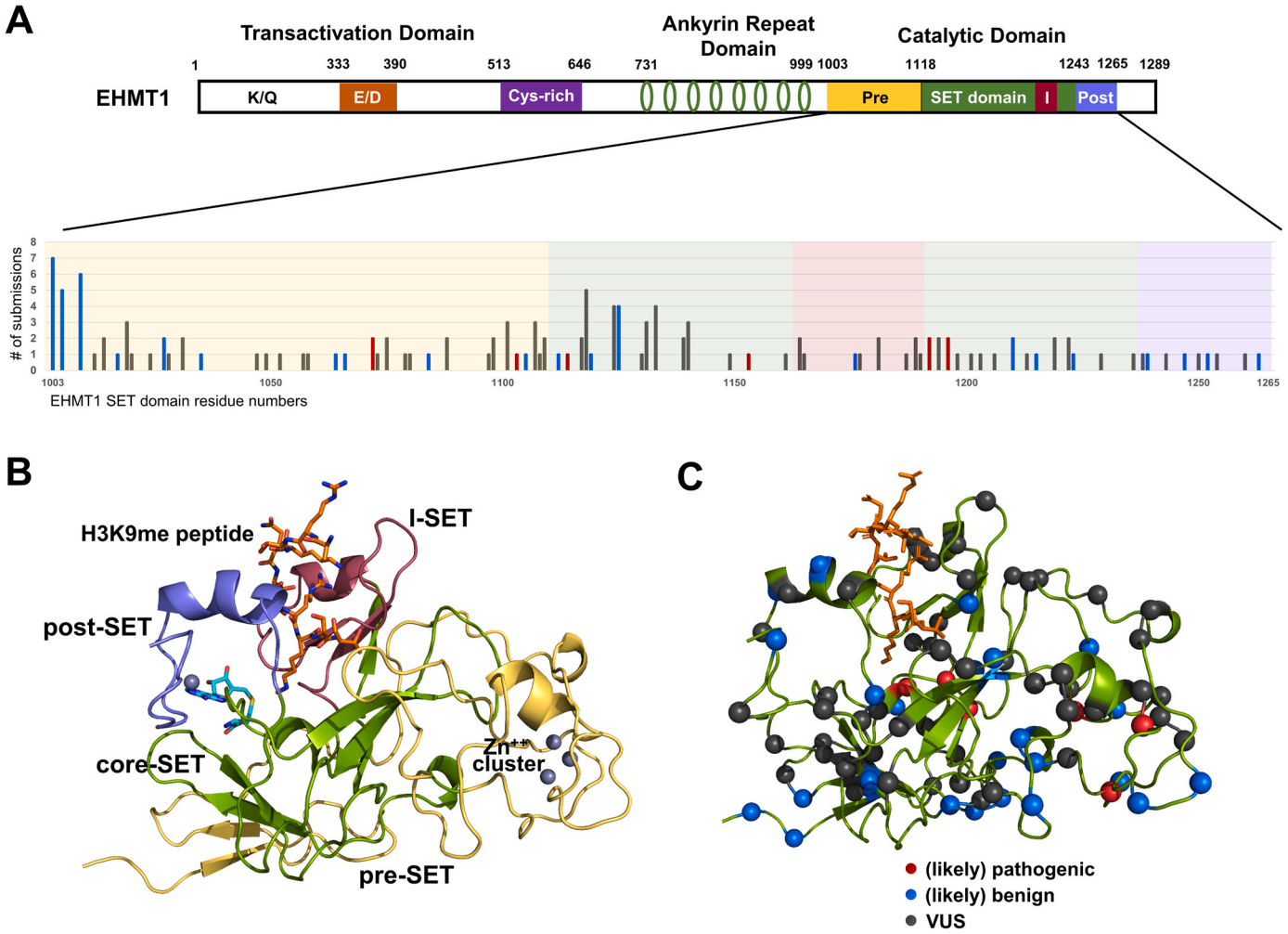

Fig. 1. Protein architecture and Kleefstra syndrome variants of EHMT1. (A) Domain structure of EHMT1 and the distribution of Kleefstra syndrome missense variants within the SET domain. In the zoomed-up window, the disease cases are tabulated by the number of independent missense mutations found on a particular residue and reported to ClinVar. The subdomains are shaded with the same colors used in Figs. 1A and 1B, and the individual bars are colored according to the current ClinVar annotations as shown in Fig. 1C. (B) Molecular structure of the EHMT1 SET domain. The same color codes for the sub-domains shown in (A) are used. The bound H3K9me peptide and the SAH cofactor are depicted as a ball-and-stick model. The structural zinc ion cluster is also indicated. (C) Current annotations of the variants under study in ClinVar and mapping of the 97 Kleefstra syndrome variants onto its molecular structure. Over 61 % of the variants (60 out of 97), 32.0 % (31 out of 97), and 6.2 % (6 out of 97) are currently classified as VUS, (likely) benign, and (likely) pathogenic, respectively. Among these, 30 variants have also been identified as cancer somatic variants. These mutations are distributed over the entire sequence and its molecular structure, with slightly higher frequencies in the pre-SET and the core SET domains.

2. Results

2.1. Defining the mutational germline landscape associated with the causality of Kleefstra syndrome

Kleefstra syndrome is a rare genetic disorder caused by mutations in EHMT1. Many of these alterations are de novo mutations [5,6]. To study the impact of SET domain missense mutations, we extracted all EHMT1 missense variants found in patients diagnosed with Kleefstra syndrome from ClinVar, the most comprehensive public archive of genomic variations and interpretations of their relationships to diseases [9]. Fig. 1A shows the distribution, frequency, and database sources for all Kleefstra syndrome variants under study. We find no apparent 'hot-spot' regions, as variants are scattered across the entire sequence and 3D structure. Slightly higher alteration frequencies are found within the pre-SET and the core-SET domains, but the differences are insignificant. Fig. 1B shows the molecular structure of the EHMT1 SET domain, and Fig. 1C shows the variant mapping onto the molecular structure and their current pathogenicity annotations in ClinVar. Although professionals routinely use these database annotations as variant interpretation guidelines, most Kleefstra syndrome variants are still classified as variants of unknown significance due to insufficient evidence and incomplete impact assessment.

Among the 97 SET domain variants, 6 are currently annotated as pathogenic or likely pathogenic, and 31 are annotated as benign or likely benign in ClinVar (Fig. 1C). Among the pathogenic variants, C1073Y and R1197W have been biochemically characterized and proven to be damaging [10] and thus serve as positive controls for this study. In addition, among the benign variants, S1004N and V1006M have relatively high allele frequencies ($> 1.0 \times 10^{-4}$) in the healthy population gnomAD database [11] and are used as potentially tolerated or neutral variants for this study. These S1004N and V1006M variants are also indicated as SNPs in the Single Nucleotide Polymorphism Database (dbSNP) and are expected to have no appreciable deleterious or pathogenic effects. Among the 97 variants, 30 have been observed somatically in human cancers [12]. While EHMT1 and EHMT2 are commonly altered in human cancers, Kleefstra syndrome variants are only found in EHMT1. Thus, this set of variants represents EHMT1-specific unique germline variants and more common disease variants.

2.2. Determining the inherent dynamic motions that characterize the molecular function of the EHMT1 SET domain

The EHMT1 SET domain is divided into sub-domains: the canonical core-SET, pre-SET, post-SET, and a small insertion within the core-SET domain architecture, termed I-SET (Fig. 1B). The post-SET and I-SET make up the substrate recognition site. In contrast, the active site is primarily located within the core-SET domain. The pre-SET region contains the structural zinc ion cluster and a dimerization interface with either EHMT1 or EHMT2 for biological homo- and hetero-dimer functional units. The MD trajectories of the wild type in complex with the SAM cofactor and the structural zinc ions show a coordinated movement with high mobility in the pre-SET region, which provides a dimerization interface. At the same time, relatively rigid motions in the active and substrate binding sites are observed (Supplementary Fig. S1 and Supplementary Movie M1). The crystal structures of the apo and H3 peptide-bound forms (PDB codes 2IGQ and 3HNA, respectively) show nearly identical conformations near the active and substrate binding sites. Thus, the histone H3 tail binds to the pre-formed stable recognition site and readily presents its H3K9me1 substrate moiety to the active site.

2.3. Comprehensive assessment of EHMT1 variants using conventional genomic tools and computational biophysics and biochemistry

We aim to combine highly correlated impact scores from each protein layer, namely sequence, structure, and dynamics. Current annotations of genomic variants are primarily based on 2D sequence conservation/residue coevolution, the physicochemical property of the substituted amino acid, and local structure considerations such as secondary structures [7,13]. Commonly used 2D sequence-based variant calling methods include SNP&Go [14], PROVEAN [15], PolyPhen2 [16], Rhapsody [17], CADD [18], and REVEL [19]. Combined annotators, such as CADD and REVEL, show better performance [20]. However, we recognize that protein function is not solely determined by its chemical composition but also by its molecular structure's spatial arrangement and dynamic nature. Therefore, we applied selective analytical measures that reflect both structural and dynamic aspects of the protein to investigate the disruptive effect of each mutation. Specifically, we conducted a series of static or dynamic structure-based analyses using either the original crystal structure (stressed) or energy-minimized structure (relaxed). Initially, we applied metrics that are universal to proteins, such as folding energy, protein stability, global/local structural perturbation, energetic frustration, and dynamics-based analyses, including root mean square deviation (RMSD), root mean square fluctuation (RMSF), a radius of gyration (Rg), and solvent accessible surface area (SASA). Subsequently, we calculated correlations among these scores to identify more functionally relevant, thus evolutionally conserved metrics for overall damaging assessment (Fig. 2). These molecular features have been essential to maintaining organismal fitness during the evolutionary selection process [21–23]. Thus, we hypothesize that integrating 2D sequence-based scores with scores from protein 3D structure and 4D time-dependent dynamic behaviors for molecular fitness will enhance the prediction power of variant interpretation as presented in our previous work [24–27]. The current study further demonstrates the importance of considering protein-specific and mechanistic-based interpretations of variants for clinical recommendations.

2.4. Analysis with the two-domain structure, congruence among individual scores, and unraveling the need to perform domain-wide analysis

We used various analytical measures that reflect both structural and dynamic aspects of the protein to probe the disruptive effect of each mutation. We initially constructed a two-domain structure containing both SET and ankyrin repeat domains by superposing the overlapping region (13 residue-long alpha-helix at the end of the ankyrin repeat domain and the beginning of the SET domain) of individual domain structures (PDB access codes 3HNA and 6BY9). Although the overlapping helix might be at a different position in each domain structure due to different crystal packing environments, the backbone torsion angles of the flexible linker (9 residues) leading the overlapping helix should have prevented the formation of unnatural or drastically shifted conformations and the energy minimization prior to MD simulation should allow recovering the low-energy native-like structure. We then performed various calculations using universal metrics from each protein layer (Fig. 2A). We used the difference values between the wild type and the variants as potential damaging scores [27]. We expected a correlation among the impact scores from all three layers of proteins, as protein sequence, structure, and dynamics are highly coupled, and their coupling strongly influences the evolutionary selection unique to each protein's molecular function [28–30]. The cross-correlation matrix calculated with the individual scores revealed that all structure-based scores showed noticeable congruence with sequence-based scores, reaffirming the interrelationship between protein structure and sequence and the effectiveness of structure-based metrics as universal metrics for all proteins (Fig. 2A). However, MD-based scores showed little congruence with the sequence-based scores, possibly due to differential sets of functionally relevant metrics for each domain with a unique function. We conducted a domain-wide impact analysis using individual domain structures to test this possibility.

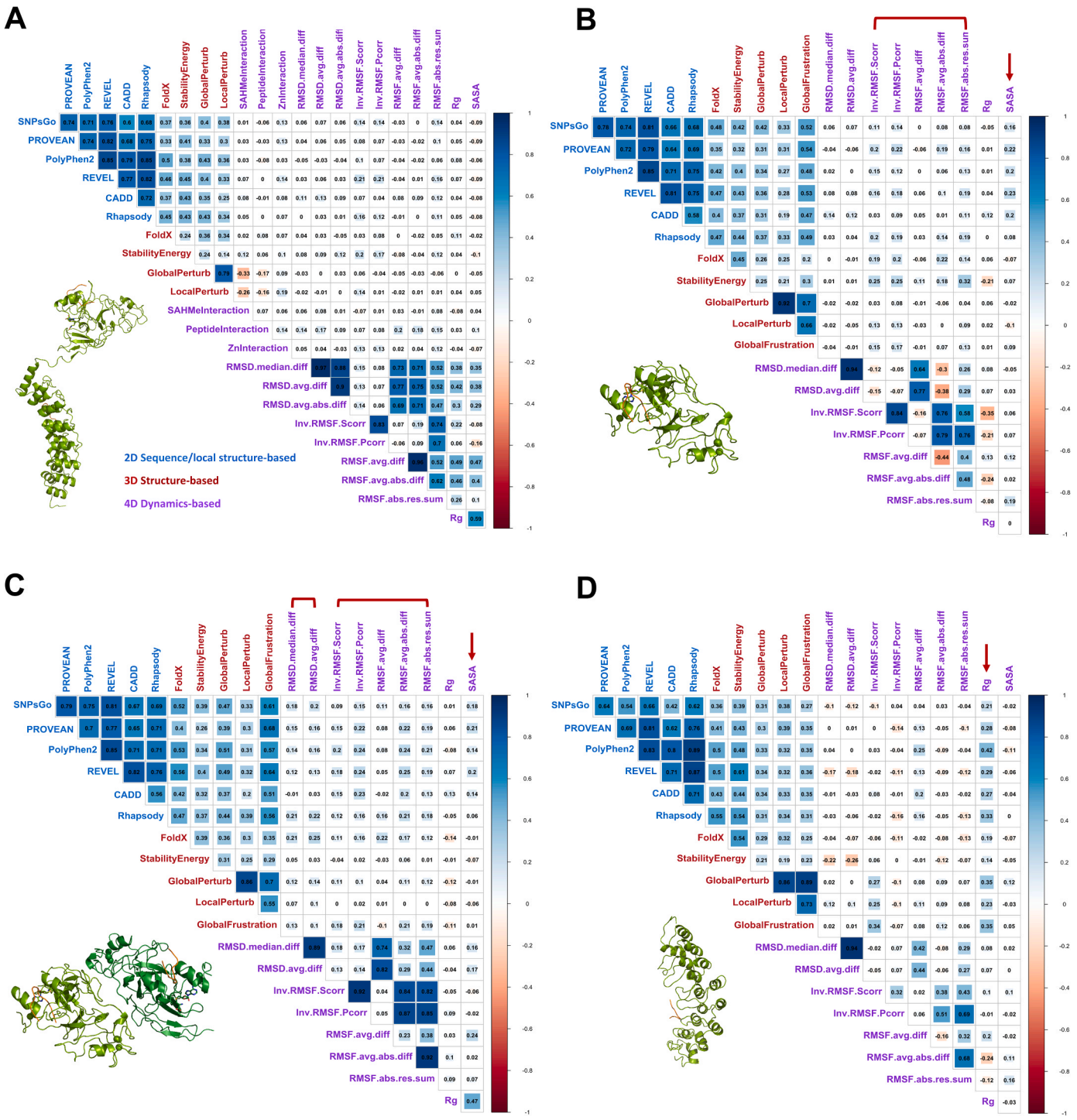

Fig. 2. Identification of congruent and domain-specific MD-based metrics by the cross-correlation matrix of the individual damaging scores. (A) Cross-correlation matrix with various measures of universal and protein-specific metrics using the two-domain structure. Although all 3D structure-based scores (red labels) show notable congruence with the 2D sequence-based scores (blue labels), nothing emerged from the 4D MD-based scores (purple labels). (B) Cross-correlation matrix with various measures of universal metrics using the SET domain-only structure. Among the 4D MD-based scores, RMSF-related scores (indicated by a square bracket on top) and SASA (indicated by an arrow) show better congruence with the sequence-based scores. (C) Cross-correlation matrix with various measures of universal metrics using the biological EHMT1:EHMT2 SET domain heterodimer structure. Among the 4D MD-based scores, in addition to RMSF-related scores and SASA, RMSD-related scores also show congruence, implying these features have been evolutionally conserved for the EHMT1 SET domain. (D) Cross-correlation matrix with various measures of universal metrics using the ankyrin repeat domain-only structure. Rg is the only congruent and functionally relevant metric among the 4D MD-based scores. A few other sporadic positive (blue) and negative (red) correlations for other MD-based scores are inconsistent and believed to be due to a higher level of background noise because of the lower number of variants in this sample.

2.5. Developing domain-specific effective and functionally relevant metrics increases the yield of discovering damaging variants

Unlike the two-domain structure analysis, when we performed the individual domain-wide analyses, positive correlations showed up for MD-based scores although the actual correlation values were lower than the structure-based scores. All structure-based scores showed significant congruence with the sequence-based scores as expected, and some dynamic-based scores showed notable agreement with other scores in a domain-specific manner. For the monomeric SET domain, RMSF-related scores and SASA showed congruence with the sequence-based scores (Fig. 2B). Despite weak correlations, these differences are quite striking, and we hypothesize that they have implications. Although the functional relevance of SASA cannot be readily explained (perhaps it is related to dimerization or some unknown protein-protein interactions), correlations with the RMSF-related scores parallel our previous findings for the Jumonji catalytic domain of KDM6A [27]. Although this needs to be tested against many different enzymes, the concerted fluctuating frequency of dynamic motions throughout the molecule might be a common property essential for all reactions to be optimally catalyzed and has been conserved [31–34], thus being an effective measure of functional disruption.

When the biological heterodimeric SET domain was used as a starting model [35], in addition to RMSF-related scores and SASA, RMSD-related scores also showed noticeable congruence (Fig. 2C). This might be related to the relative orientation between the monomers, which can play a critical functional role (Supplementary Fig. S2). However, dimerization interaction energies show weaker correlations with the sequence-based scores (Supplementary Fig. S3). On the other hand, for the alpha-solenoid ankyrin repeat scaffolding domain, only Rg was a congruent and effective metric for functional disruption (Fig. 2D). For EHMT1, each domain has a different set of functionally relevant and more effective metrics for functional disruption. When scored collectively, the true signals can be canceled out, producing low correlations with other scores (Fig. 2A). As a result, the MD-based damaging scores from the two-domain model can become less reliable. Thus, we used the individual domain structures to complete domain-specific and comprehensive molecular fitness analyses. In this article, we present the data with the catalytic SET domain; the results with the ankyrin repeat domain will be published later with validation data. The entire molecular fitness scores for the SET domain are provided in Supplementary Table S1, and their plots against the variants for each metric are shown in Supplementary Fig. S4.

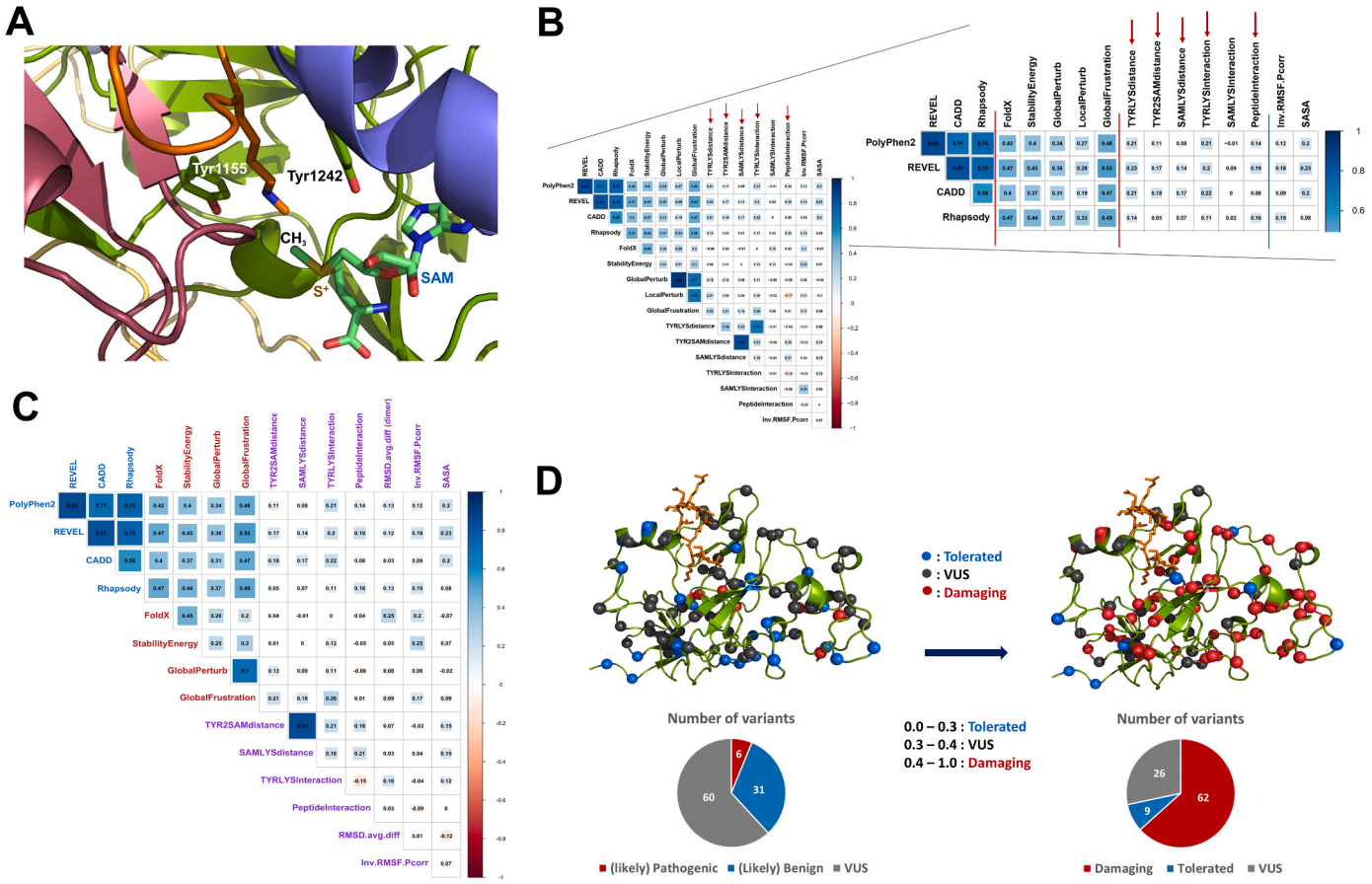

Fig. 3. Identification of additional domain function-specific MD-based metrics, the final set of metrics used for overall impact scoring, and reclassification of the variants. (A) Key functional elements in the active site. In addition to the SAM cofactor, two tyrosine residues (Tyr 1155 and Tyr 1242) play critical roles in catalysis. (B) Cross-correlation matrix with additional catalysis-related metrics and the universal MD-based metrics previously identified such as RMSF and SASA (separated by the cyan bar among the MD-based metrics). Noticeably congruent and functionally relevant domain-specific metrics are indicated by red arrows. Tyr and Tyr2 refer to the aligning Tyr1155 and the catalytic Tyr1242 residues, respectively. (C) Cross-correlation matrix of the scores from the finally chosen metrics for meta-score calculations that are concordant and functionally relevant, thus have been evolutionally conserved. (D) Mapping of pre- and post-classified EHMT1 Kleefstra syndrome variants. Meta-scoring reveals that the damaging variants (red) are concentrated near the substrate binding site, active site, and pre-SET region which contains the structural zinc-ion cluster and the dimerization interface, while the tolerated variants (blue) are primarily found in the periphery or surface of the molecule. Pie charts at the bottom show the numbers of the pre- and post-classified variants in each category.

2.6. Parametrizing SET domain-specific metrics extends the predictive power of evaluating damaging variants in enzymatic domains

The SET domain is well-known for its catalytic mechanism involving critical functional roles of the SAM cofactor and two tyrosine residues [36,37]. Tyr1155 aligns the substrate lysine for the methyl transfer reaction, while Tyr1242 enhances the electrophilicity of the departing methyl group of the SAM cofactor (Fig. 3A). Optimal enzymatic activity relies on local geometry and mutual interactions, which can be affected by disease-causing mutations. Thus, we measured time-dependent interaction energies to assess potential damaging impacts and monitored the distances between critical functional elements from the molecular dynamics (MD) trajectories. We also calculated the differences between the wild type and the variants for the peptide interaction energy (substrate binding). We observed congruence between the sequence-based scores and all essential interaction- and distance-related scores, except for the SAM cofactor and the substrate target Lys interactions, when we calculated cross-correlations between these values and other scores (Fig. 3B). Thus, in the overall impact scoring, we included these domain-specific metrics in addition to the congruent universal metrics such as RMSF and SASA (Fig. 3C). This confirmed the functional relevance of these metrics and supported their inclusion in the overall impact scores.

2.7. Data integration and modeling allow meta-scoring and re-classification of EHMT1 SET domain Kleefstra syndrome variants

We selected four structure-based and seven dynamics-based metrics that show congruence with sequence-based scores and combined them with four sequence-based scores to extend the predictive power of evaluating damaging variants in enzymatic domains (Fig. 3C). These metrics were used to compute an overall score for each variant, which can be numerically represented for practical use by clinicians and geneticists. However, calculating the final scores simply by summing up the individual scores is inappropriate, as the measurements for individual metrics are given in different units and have distinct ranges. Therefore, we chose to use Z-score conversion and scaling to transform the individual scores into a zero to one range, commonly used by many sequence-based tools [38]. We then averaged them for the final scores (Supplementary Table S1).

We used suggested thresholds for each prediction tool as guidelines to classify the variants. We re-classified the variants into three groups: tolerated (0–0.3), uncertain (0.3–0.4), and damaging (0.4–1.0) based on their overall damaging scores. The re-classified variants are shown in Fig. 3D. This results in a balanced number of variants in each category. Out of 97 variants evaluated, 62 (63.9 %) were classified as damaging, 9 (9.3 %) were classified as tolerated, and 26 (26.8 %) remained as variants of uncertain significance (VUS). However, this significantly improved from the 60/97 (61.9 %) VUS identified in the current ClinVar annotations. The damaging variants are mostly found near the functional regions, while the tolerated variants are located on the periphery or molecular surface. Our analysis showed that the annotations of currently classified VUS have been improved using our method, and our comprehensive structural genomics approach enhanced the prediction power of genomic variant interpretation.

2.8. A molecular biophysics classification of variants and EHMT1/2 paralog analysis extends information on damaging effects on EHMT1 and generalizes results to related proteins

We further classified the damaging variants into structural (SV), dynamic (DV), and structural & dynamic variants (SDV) to provide mechanistic interpretations [27]. We calculated molecular fitness scores by considering only structure- and dynamics-based scores, as shown in Supplementary Table S1. These molecular fitness evaluations revealed that 13 variants are expected to disrupt at least one of the structural features. In comparison, 51 variants are expected to disrupt at least one of the dynamic features (Fig. 4A). Among these, 11 variants disrupted both the protein's structural and dynamic properties. The C1073Y and R1197W variants, previously characterized and proven to be damaging to protein function, are predicted to be damaging by our analysis (Fig. 4A, left panel). In contrast, our analysis predicts two other variants from gnomAD with a relatively high allele frequency $>1.0 \times 10^{-4}$ in general populations (S1004N and V1006M variants) to be tolerated by our analysis (Fig. 4A, right panel). All previously annotated damaging variants are expected to be damaging by our analysis (Fig. 4B, left pie chart). However, among the previous benign annotators, only 4 out of 31 are expected to be tolerated by our analysis, and 18 are expected to be damaging while 9 variants now belong to the VUS group (right pie chart). Thus, our results suggest that current annotations in the database tend to underestimate the damaging impact of genomic variants. The annotations of the currently classified VUS variants have also been improved using our method, and 38 out of 60 are expected to be damaging while 5 are expected to be tolerated (middle pie chart). Our analyses provide evidence that integrating 2D sequence-based scores with the scores from the protein 3D structure and 4D time-dependent dynamic behaviors for molecular fitness can enhance the prediction power of variant interpretation and provide potential molecular mechanisms for functional disruption.

Finally, we evaluated the results against the sequence conservation between EHMT1 and EHMT2. Our data indicated that most damaging variants (57/62, 91.9 %) are found on canonical residues that are highly conserved for structural and functional reasons while most tolerated ones (8/9, 88.9 %) are found in varying residues (Fig. 4C). None of the tolerated germline variants are represented in the cancer somatic variants. Cancer somatic variants are found throughout the sequence including many varying residues, some of which might represent polymorphisms and be tolerated. Overall, our data showed that the damaging consequences of these variants are well represented in the human disease genomic landscape. This sequence information and mechanistic-based structural bioinformatics have the potential to provide better diagnosis, risk assessment, and clinical guidelines for observed variants within individualized medicine.

3. Discussion

Proteins perform vital functions by being made up of amino acids that fold into a unique 3D structure. The effect of genetic variation on protein structure and function can be dramatic, with non-synonymous SNPs being the most common DNA sequence variation associated with human diseases [39]. Missense mutational effects can alter protein dynamics and cause human disease. Accurate evaluation of these effects can provide information on residue-specific roles in protein structure/function and dysfunction in the disease state. The current study advances the field of rare diseases, particularly Kleefstra syndrome, by implementing a computational biophysics approach and contrasting it with previous tools recommended by established guidelines.

The sequence-structure-function relationship has been established for all proteins, but molecular dynamics still need to be fully explored in genomic variation interpretation. To improve the assessment of genomic variations, we implemented a comprehensive computational approach incorporating multiple mechanism-based aspects of the protein sequence, structure, and dynamics of EHMT1 for its mutational impact assessment. Structurally coordinated dynamics play an essential role in substrate binding and undergoing allosteric transitions while maintaining the native fold in catalytic enzymes [40,41]. Thus, dynamics-related protein-specific metrics can be reliable indicators of any protein function and dysfunction caused by disease mutations. Once identified, these metrics can be parameterized for each protein and domain. We analyzed each variant independently to show how curated missense variants may affect EHMT1 enzymatic activities. The selected metrics used in the current study served as effective measures of

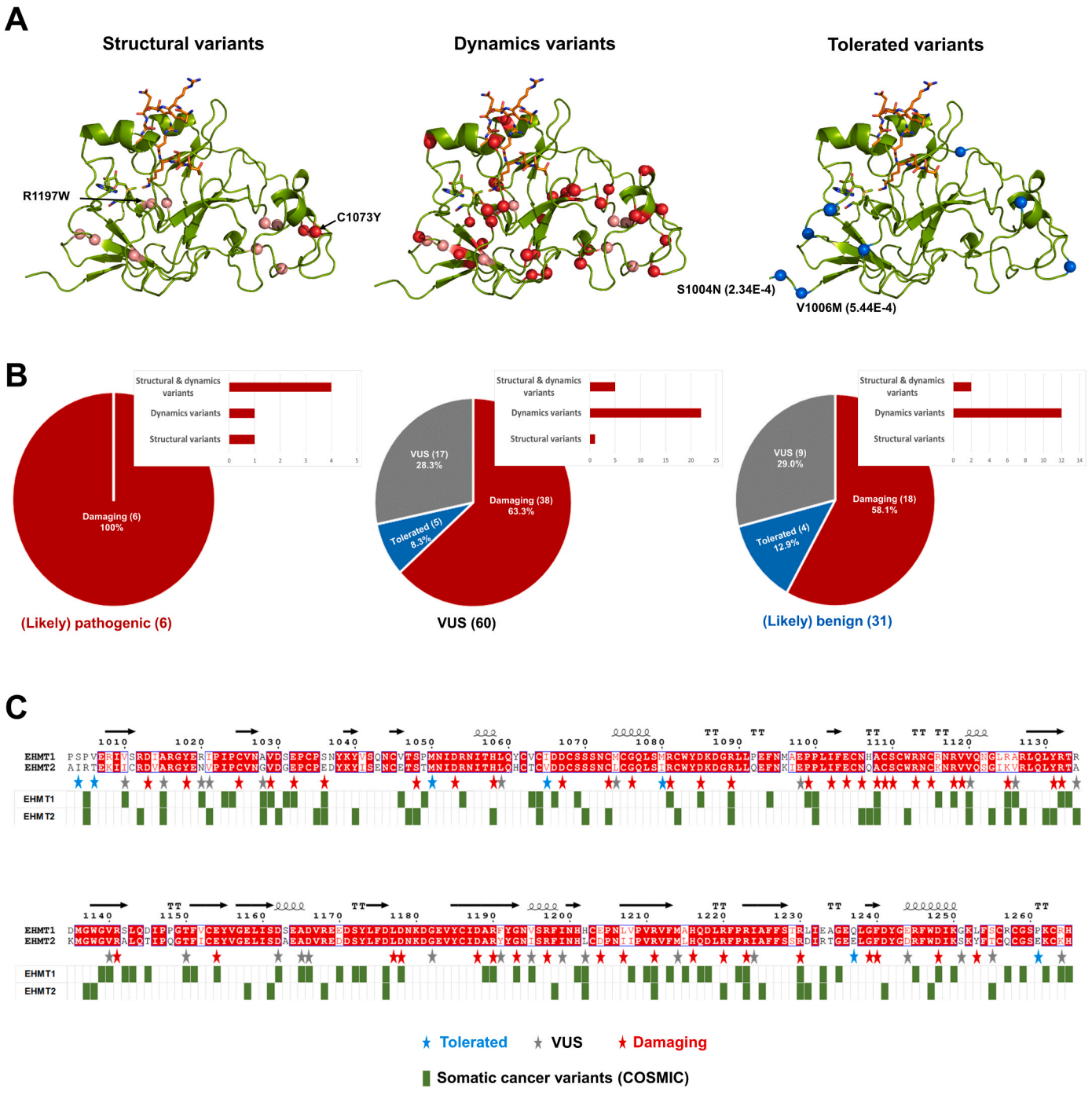

Fig. 4. Further classification of the variants, comparison with pre-classification, and the EHMT1/2 paralog analysis. (A) Additional classification of the variants into structural variants (SV), dynamic variants (DV), and structural & dynamic variants (SDV), and mapping onto the molecular structure. SV or DV are indicated as red balls, SDVs are indicated in pink, and the tolerated variants are shown in blue. Previously characterized damaging variants (C1073Y and R1197W) and higher allele frequency variants among the healthy populations (S1004N and V1006M) are also indicated. (B) Comparison of pre-classification (ClinVar annotations) and post-classification by our molecular fitness analysis for each pre-classified group. We compared the two classification results using a pie chart that indicates damaging versus tolerated for our new classification results for each of the three pre-classification categories. The inset bar chart shows the balance among the three damaging sub-categories. The damaging variants influenced by only 2D-based scores are not considered in the inset bar charts. These mechanism-based interpretations should help resolve the conflicting variants (middle) and provide enhanced interpretations for all variants. The numbers of the variants in each group are indicated in parentheses. (C) Comparison of post-classification and EHMT paralog annotation analysis. Sequence alignment of human EHMT1 and EHMT2 with indications of conserved residues (red boxes) and revised annotations (stars below). More than one variant is found on some of these residues. Cancer somatic variants in both proteins are also indicated at the bottom (green boxes). Residue numbers are based on EHMT1. Secondary structure elements are shown at the top of each sequence alignment. Positions of the re-classified variants of the current study are indicated by red (damaging), VUS (black), and blue (tolerated) stars at the bottom of aligned sequences. All germline variants on the conserved residues are predicted to be damaging by our comprehensive analysis, and none of the tolerated germline variants are represented in the cancer somatic variants.

damaging impacts. The cross-correlation matrix of the individual scores was used to select more functionally relevant and effective metrics for each protein. Our findings affirm that the protein sequence-structure-dynamics-function relationships, molecular dynamic properties, and molecular structures have been conserved throughout evolution.

We hypothesize that universal structure-based metrics, such as folding/stability, global/local structural perturbation, and global/local energetic frustration, can be applied to all proteins and become part of the standard procedure for clinical functional impact analysis. On the other hand, MD-based metrics are not equally effective, and more congruent and functionally relevant MD-based metrics need to be identified for each protein. Moreover, a domain-wide analysis should be considered, especially when individual domains do not make any physical contact, because individual domains are independent folding units and functional modules.

Many proteins consist of several domains, and the same domain may appear in various proteins. Because members of the same domain family likely share the same evolutionary origin and perform similar molecular functions, the same set of effective MD-based metrics can be applied to the same domain family members. In many cases, long stretches of disordered regions connect individual domains, and only individual domain structural information is available. Even when multi-domain structures are available, individual domains can be isolated during data analysis and used for domain-specific analysis to gain more domain function-specific impact analysis. Subsequent collective analysis on multi-domain or functional oligomeric structures can provide additional mechanistic information, such as interdomain communication and cooperative functionality.

The current study lacks a distinct training set due to the rarity of the disease and the very few genotype-phenotype relationship studies available for EHMT1, and all ClinVar variants were treated as a test set. However, the results with our control variants and the unity of the metrics with known functional relevance, such as the substrate (H3 peptide) interactions and the available roles of the tyrosine residues in the active site, all support the effectiveness of our approach. We firmly believe that the structure- and dynamics-based analyses will critically augment the prediction power beyond sequence-based benchmarking tools and improve the overall impact assessment. For future improvement, more extensive studies such as different simulation conditions and time scales will be considered. Additional impact assessments such as protein expression, impact on RNA structures, translocation, protein–-protein interactions, post-translational modifications, etc. will be included in the workflow.

In conclusion, the current work provides important molecular-level insights into functional disruption by Kleefstra syndrome variants. Our data indicate that damaging variants of EHMT1 display mechanistic disruptions at either a structural or dynamics level or both, mainly concentrated around functional regions such as the active site and the substrate binding interface and the pre-SET region that contains the structural zinc ion cluster and the dimerization interface. On the other hand, tolerated variants are mostly found on the periphery or on the molecular surface, whose sequences are varied between EHMT1 and EHMT2. These findings should apply to not only EHMTs but also related SET domain-containing methyltransferases. Extended studies that use sequence paralogs and molecular dynamics will help validate our findings and improve the predictive value of these mechanistic-based comprehensive approaches. Furthermore, this comprehensive impact analysis should help annotate the pathogenicity of many different proteins that can be curated into the public archives of human genomic variations for clinical applications.

4. Materials and methods

4.1. The extraction of KS variants from the public archive and the selection of control variants

KS-associated variants were identified from ClinVar, the most comprehensive public archive of genomic variations and interpretations of their relationships to diseases [9]. All 97 EHMT1 missense variants found in patients diagnosed with KS and reported in ClinVar at the commencement of this study were extracted. These variants were referenced against cancer somatic variants curated in the database such as the Catalog of Somatic Mutations in Cancer (COSMIC) [12] and TCGA [42], which revealed that 30 out of 97 KS variants are also reported as cancer somatic variants. An independent literature search identified the C1073Y and R1197W variants as experimentally proven damaging variants and thus serve as positive controls [10]. In addition, among the benign variants, the S1004N and V1006M variants are also indicated as SNPs in the Single Nucleotide Polymorphism Database (dbSNP) and have relatively high allele frequencies in the general population gnomAD database [11], thus serving as neutral variants or negative controls.

4.2. Preparation of the initial structures

4.2.1. We constructed the two-domain structure containing

both the SET and the ankyrin repeat domains by superposing the overlapping helix (982–998) of the individual domain structures (PDB access codes 3HNA and 6BY9) after considering the asymmetric contents of the crystal lattice. For the SET domain-only structure, we used the mono-methylated H3K9 peptide-bound form of the high-resolution (1.5 Å) crystal structure (PDB access code 3HNA). The cofactor SAH was replaced with SAM to constitute the active form of the enzyme, and the seven missing residues in the flexible loop (965–971) were built using the Modeller program [43]. The EHMT1-EHMT2 SET domain-only heterodimer was prepared by replacing its heterodimer partner from either homodimer structures (PDB access codes 3HNA or 5JJ0), which display a nearly identical dimerization binding mode. For missense variant analysis using these structures, substitutions were made within the Discovery Studio suite version 21.1 (Dassault Systèmes BIOVIA) by mutating the corresponding residue and selecting the side chain rotamer causing the least steric hindrance with the surrounding residues.

4.3. Protein folding energy and stability calculation

We assessed the stability of the mutated protein by the variant-induced changes in folding energy ($\Delta\Delta G_{fold}$) using FoldX [44] and the Discovery Studio suite. We used the energy-minimized mutant structures for these calculations. In Discovery Studio, shifted amounts of protein stability (free-energy difference between folded and unfolded states) due to mutations were calculated at pH 7.4 using the energy-minimized wild type structure and introducing each substitution for calculation. After the preparation phase, the initial structures of the wild type and the generated mutants were subjected to a two-stage minimization process before energy calculation. The predicted $\Delta\Delta G$s, using both programs, are in good agreement (Supplementary Table S1 and Supplementary Fig. S4A-B).

4.4. Global and local structure perturbation measurements

We measured the positional displacement of backbone atoms between the entire catalytic domains of wild type and mutant (global) or only the atoms near the residue of interest between them (local). For local structure perturbation from the energy-minimized structures, any residues within a 10 Å radius of the mutation site were selected using PyMOL (Schrödinger, LLC) and calculated for least-squared RMSD of the backbone atoms between the wild type and the mutant using Coot [45].

For global structure perturbation, entire backbone atoms were used for RMSD calculation between the structures.

4.5. Frustration index calculation

The energy landscape of protein molecules can affect their biological behaviors. To evaluate how energy is distributed in protein structures and how mutations or conformational changes shift the energy distributions, we measured the differences in energetic frustration in protein molecules using the Protein Frustratometer server [46]. The *frustration index* measures how favorable a particular contact is relative to all possible contacts in that location. Sites of high local frustration often indicate biologically important regions such as binding or allosteric sites. The shift amounts were calculated by measuring the differences in the frustration indexes between the wild type and mutant residues in both directions and summing up the differences. We tested the changes in either global (cumulative) or local frustration indexes as a means of damaging impact scoring. We discovered that the global changes show better correlations with the sequence-based scores, which aligns with the multiple binding platforms used by this domain for various protein-protein interactions. Therefore, global changes were used for the overall impact scoring of the variants in the ankyrin repeat domain.

4.6. Molecular simulations

MD simulations were performed using the CHARMm36 all-atom-force field [36] implemented in Discovery Studio with a 2 fs time step. A simplified distance-dependent implicit solvent environment was used with a dielectric constant of 80 and a pH of 7.4. All MD simulations were conducted using periodic boundary conditions. Models were energy minimized for 5000 steps using the steepest descent followed by 5000 steps of the conjugate gradient to relax the protein structure obtained under the stressed crystal environment. Each system of 10 replicates of wild type and each variant was independently heated to 300 K over 200 ps and equilibrated for 500 ps, followed by ten ns production simulation under the NPT ensemble (100 ns total). Conformers were recorded every 10 ps to give 1000 frames for analysis per each mutation. This timescale is sufficient for side chain rearrangements in the protein's native state and to facilitate local conformational changes. The total energy plots of the trajectories indicate that the systems can reach near equilibrium towards the end of the simulation. For final data analysis, one or two outliers (in some cases none) from each data set of 10 replicates that deviate from the rest in RMSD plots and might represent the minor and rarer form of conformations (altogether 14 % of the entire data) were excluded from averaging, and only the last 500 frames that have reached the near minimum total energy state were used. From a 10 ns MD simulation, trajectory files were analyzed for structural impact by root mean squared deviation (RMSD), root mean square fluctuation (RMSF), and other measures such as time-dependent molecular interactions, a radius of gyration (Rg), and solvent-accessible surface area (SASA). Trajectories were aligned to the initial WT conformation before analysis. RMSD and RMSF values were calculated at the residue level for all atoms using the tools available within Discovery Studio and the algorithms implemented in Microsoft Excel. Further analyses were conducted in the R programming language [47], leveraging the bio3d package [48]. Molecular visualizations were generated using PyMOL.

4.7. Time-dependent interaction energy calculation and distance monitoring

Molecular interaction-free energies were measured using Discovery Studio. This was done using the MD simulation trajectories and selecting the protein and the interaction groups of interest. Non-bonded interactions were monitored, and dynamic interaction energies (van der Waals and electrostatic energies) were calculated using the CHARMm36 force field and the implicit distance-dependent dielectric solvent model. Distance monitoring of the key catalytic components was also done within Discovery Studio by selecting those atoms of interest. These measurements were made for all 10 replicates and averaged for comparison with the wild type.

4.8. Overall impact classification of the variant

We used a cross-correlation matrix among the scores as guidance to choose more effective and functionally relevant metrics for integration (Supplementary Text S1). For overall impact scoring, we Z-score scaled the individual scores onto a zero to one scale, commonly used by many sequence-based tools, and averaged them for the final scores [27]. Finally, for variant classification, we used the suggested thresholds for sequence-based prediction tools for overall impact scoring as guidelines and re-classified the variants based on the meta-scores (0–0.3: tolerated, 0.3–0.4: uncertain, and 0.4–1.0: damaging). This results in a balanced number of variants in each category. As a result, known damaging variants and gnomAD healthy population variants belong to the respective expected categories. Likewise, the proper threshold values were chosen for molecular fitness scores (without the sequence-based scores) based on the suggested values for sequence-based prediction tools. Any variants predicted to be damaging in either molecular aspect (structure or dynamics), yet the overall meta-scores below 0.4 have been assigned as VUS. Similarly, any variants that fall below the threshold value in either aspect, yet collectively the overall meta-scores exceed 0.4 have been assigned as damaging (but not further classified as either a structural or dynamic variant).

CRediT authorship contribution statement

RU devised the project, the main conceptual ideas, and the proof outline. RU, MTZ, SDJ, and YC designed the computational framework and analyzed the data. DRJ, BCS, BFV, AJM, and GL discussed the results, provided critical feedback, and helped shape the research and analysis. YC, MTZ, and RU were major contributors to writing the manuscript. All authors read and approved the final manuscript.

Declaration of Competing Interest

The authors declare that they have no competing interests.

Data Availability

MD Simulation data is available upon request.

Acknowledgments and funding information

This work was funded by the Advancing a Healthier Wisconsin Endowment and the 501c3 charitable organization Harmony 4 Hope to the Precision Medicine Simulation Unit of the Mellowes Center for Genomic Sciences and Precision Medicine at the Medical College of Wisconsin (to RU). This work was also supported in part by NIH grants R35GM128840 (to BCS) and R01DK052913 (to RU and GL).

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2023.10.022.

References

[1] Fahrner JA, Bjornsson HT. Mendelian disorders of the epigenetic machinery: postnatal malleability and therapeutic prospects. Hum Mol Genet 2019;28(R2): R254–64.
[2] Shinkai Y, Tachibana M. H3K9 methyltransferase G9a and the related molecule GLP. Genes Dev 2011;25(8):781–8.

[3] Rahman Z, Bazaz MR, Devabattula G, Khan MA, Godugu C. Targeting H3K9 methyltransferase G9a and its related molecule GLP as a potential therapeutic strategy for cancer. J Biochem Mol Toxicol 2021;35(3):e22674.
[4] Zhang T, Cooper S, Brockdorff N. The interplay of histone modifications - writers that read. EMBO Rep 2015;16(11):1467–81.
[5] Willemsen MH, Vulto-van Silfhout AT, Nillesen WM, Wissink-Lindhout WM, van Bokhoven H, Philip N, et al. Update on Kleefstra syndrome. Mol Syndr 2012;2(3–5):202–12.
[6] Ciaccio C, Scuvera G, Tucci A, Gentilin B, Baccarin M, Marchisio P, et al. New insights into Kleefstra syndrome: report of two novel cases with previously unreported features and literature review. Cytogenet Genome Res 2018;156(3):127–33.
[7] Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med 2015;17(5):405–24.
[8] Rehm HL, Fowler DM. Keeping up with the genomes: scaling genomic variant interpretation. Genome Med 2019;12(1):5.
[9] Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res 2014;42:D980–5 (Database Issue).
[10] Yamada A, Shimura C, Shinkai Y. Biochemical validation of EHMT1 missense mutations in Kleefstra syndrome. J Hum Genet 2018;63(5):555–62.
[11] Gudmundsson S, Singer-Berk M, Watts NA, Phu W, Goodrich JK, Solomonson M, et al. Variant interpretation using population databases: lessons from gnomAD. Hum Mutat 2022;43(8):1012–30.
[12] Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, et al. COSMIC: the catalogue of somatic mutations in cancer. Nucleic Acids Res 2019;47(D1):D941–7.
[13] Li MM, Datto M, Duncavage EJ, Kulkarni S, Lindeman NI, Roy S, et al. Standards and guidelines for the interpretation and reporting of sequence variants in cancer: a joint consensus recommendation of the association for molecular pathology, American Society of Clinical Oncology, and College of American Pathologists. J Mol Diagn 2017;19(1):4–23.
[14] Capriotti E, Calabrese R, Fariselli P, Martelli PL, Altman RB, Casadio R. WS-SNPs&GO: a web server for predicting the deleterious effect of human protein variants using functional annotation. BMC Genom 2013;14(Suppl 3):S6 (Suppl 3).
[15] Choi Y, Chan AP. PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. Bioinformatics 2015;31(16):2745–7.
[16] Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nat Methods 2010;7(4):248–9.
[17] Ponzoni L, Penaherrera DA, Oltvai ZN, Bahar I. Rhapsody: predicting the pathogenicity of human missense variants. Bioinformatics 2020;36(10):3084–92.
[18] Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res 2019;47(D1):D886–94.
[19] Ioannidis NM, Rothstein JH, Pejaver V, Middha S, McDonnell SK, Baheti S, et al. REVEL: an ensemble method for predicting the pathogenicity of rare missense variants. Am J Hum Genet 2016;99(4):877–85.
[20] Wang D, Li J, Wang Y, Wang E. A comparison on predicting functional impact of genomic variants. NAR Genom Bioinform 2022;4(1):lqab122.
[21] Romero PA, Arnold FH. Exploring protein fitness landscapes by directed evolution. Nat Rev Mol Cell Biol 2009;10(12):866–76.
[22] Tomatis PE, Fabiane SM, Simona F, Carloni P, Sutton BJ, Vila AJ. Adaptive protein evolution grants organismal fitness by improving catalysis and flexibility. Proc Natl Acad Sci U S A 2008;105(52):20605–10.
[23] Meini MR, Tomatis PE, Weinreich DM, Vila AJ. Quantitative description of a protein fitness landscape based on molecular features. Mol Biol Evol 2015;32(7):1774–87.
[24] Dong Z, Zhou H, Tao P. Combining protein sequence, structure, and dynamics: a novel approach for functional evolution analysis of PAS domain superfamily. Protein Sci 2018;27(2):421–30.
[25] Ponzoni L, Bahar I. Structural dynamics is a determinant of the functional significance of missense variants. Proc Natl Acad Sci U S A 2018;115(16):4164–9.
[26] Tripathi S, Dsouza NR, Urrutia R, Zimmermann MT. Structural bioinformatics enhances mechanistic interpretation of genomic variation, demonstrated through the analyses of 935 distinct RAS family mutations. Bioinformatics 2020.
[27] Chi YI, Stodola TJ, De Assuncao TM, Leverence EN, Smith BC, Volkman BF, et al. Structural bioinformatics enhances the interpretation of somatic mutations in KDM6A found in human cancers. Comput Struct Biotechnol J 2022;20:2200–11.
[28] Narayanan C, Bernard DN, Bafna K, Gagne D, Chennubhotla CS, Doucet N, et al. Conservation of dynamics associated with biological function in an enzyme superfamily. Structure 2018;26(3):426–36. e3.
[29] Liu Y, Bahar I. Sequence evolution correlates with structural dynamics. Mol Biol Evol 2012;29(9):2253–63.
[30] Maguid S, Fernandez-Alberti S, Echave J. Evolutionary conservation of protein vibrational dynamics. Gene 2008;422(1–2):7–13.
[31] Warshel A. Dynamics of enzymatic reactions. Proc Natl Acad Sci U S A 1984;81(2):444–8.
[32] Boekelheide N, Salomon-Ferrer R, Miller 3rd TF. Dynamics and dissipation in enzyme catalysis. Proc Natl Acad Sci U S A 2011;108(39):16159–63.
[33] Gagne D, Doucet N. Structural and functional importance of local and global conformational fluctuations in the RNase A superfamily. FEBS J 2013;280(22):5596–607.
[34] Warshel A, Bora RP. Perspective: defining and quantifying the role of dynamics in enzyme catalysis. J Chem Phys 2016;144(18):180901.
[35] Sanchez NA, Kallweit LM, Trnka MJ, Clemmer CL, Al-Sady B. Heterodimerization of H3K9 histone methyltransferases G9a and GLP activates methyl reading and writing capabilities. J Biol Chem 2021;297(5):101276.
[36] Qian C, Zhou MM. SET domain protein lysine methyltransferases: structure, specificity and catalysis. Cell Mol Life Sci 2006;63(23):2755–63.
[37] Schapira M. Structural chemistry of human SET domain protein methyltransferases. Curr Chem Genom 2011;5(Suppl 1):85–94.
[38] Wu Y, Liu H, Li R, Sun S, Weile J, Roth FP. Improved pathogenicity prediction for rare human missense variants. Am J Hum Genet 2021;108(12):2389.
[39] Shastry BS. SNPs in disease gene mapping, medicinal drug development and evolution. J Hum Genet 2007;52(11):871–80.
[40] Ramanathan A, Agarwal PK. Evolutionarily conserved linkage between enzyme fold, flexibility, and catalysis. PLoS Biol 2011;9(11):e1001193.
[41] Vendruscolo M, Dobson CM. Structural biology. Dynamic visions of enzymatic reactions. Science 2006;313(5793):1586–7.
[42] Grossman RL, Heath AP, Ferretti V, Varmus HE, Lowy DR, Kibbe WA, et al. Toward a shared vision for cancer genomic data. N Engl J Med 2016;375(12):1109–12.
[43] Webb B, Sali A. Protein structure modeling with MODELLER. Methods Mol Biol 2017;1654:39–54.
[44] Guerois R, Nielsen JE, Serrano L. Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. J Mol Biol 2002;320(2):369–87.
[45] Emsley P, Cowtan K. Coot: model-building tools for molecular graphics. Acta Crystallogr D Biol Crystallogr 2004;60(Pt 12 Pt 1):2126–32.
[46] Parra RG, Schafer NP, Radusky LG, Tsai MY, Guzovsky AB, Wolynes PG, et al. Protein frustratometer 2: a tool to localize energetic frustration in protein molecules, now with electrostatics. Nucleic Acids Res 2016;44(W1):W356–60.
[47] Team R.C.R.: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL ⟨https://www.R-project.org/⟩. 2020.
[48] Grant BJ, Rodrigues AP, ElSawy KM, McCammon JA, Caves LS. Bio3d: an R package for the comparative analysis of protein structures. Bioinformatics 2006;22(21):2695–6.