NAR Genomics and Bioinformatics, 2025, **7**, lqaf047
https://doi.org/10.1093/nargab/lqaf047
Methods Article

OXFORD

Detection of protein structural hotspots using AI distillation and explainability: application to the DAX-1 protein

Noé Dumas [1,*], Geoffrey Portelli [1], Yang Ji [1], Florent Dupont [1], Mehdi Jendoubi [1], Enzo Lalli [2,3,4,*]

[1] Thales SA, Thales Services Numériques, 06560 Valbonne—Sophia Antipolis, France
[2] Centre National de la Recherche Scientifique, Institut de Pharmacologie Moléculaire et Cellulaire, 06560 Valbonne—Sophia Antipolis, France
[3] Institut national de la santé et de la recherche médicale, Institut de Pharmacologie Moléculaire et Cellulaire, 06560 Valbonne—Sophia Antipolis, France
[4] Université Côte d'Azur, Institut de Pharmacologie Moléculaire et Cellulaire, 06560 Valbonne—Sophia Antipolis, France

[*] To whom correspondence should be addressed. Email: ninino@ipmc.cnrs.fr
Correspondence may also be addressed to Noé Dumas. Email: noe.dumas@thalesgroup.com

Abstract

AlphaMissense is a valuable resource for discerning important functional regions within proteins, providing pathogenicity heatmaps that highlight the pathogenic risk of specific mutations along the protein sequence. However, due to protein folding and long-range interactions, the actual structural alterations with functional implications may be occurring at a distance from the mutation site. As a result, the identification of the most sensitive structural regions for protein function may be hampered by the presence of mutations that indirectly affect the critical regions from a distance. In this study, we illustrate how the use of AlphaMissense predictions to train an XGBoost regression model on structural features extracted from the structures of protein variants predicted by OmegaFold enables the definition of a new explainability metric: a residue-based importance score that highlights the most critical structural domains within a protein sequence. To verify the accuracy of our approach, we applied it to the extensively studied protein DAX-1 and successfully identified critical structural domains. Notably, as this score only requires knowledge of the protein's amino acid sequence, it is valuable in guiding experimental investigations aimed at discovering functionally crucial regions in proteins that have been poorly characterized.

Introduction

The accurate identification of the most critical regions of a protein for the characterization of its structure and function is invaluable for comprehending its underlying mechanism of action and for directing the development of pharmaceuticals that modulate its activity.

Numerous methods have been created to identify these regions, often referred to as hotspots [1–5]. Most of these techniques focus on the detection of binding hotspots within protein–protein interfaces (PPIs). These hotspots are specific regions on the protein surface, marked by a high concentration of residues that significantly contribute to the overall binding energy.

Even though PPIs are attractive for drug development due to their evident functional significance, they often consist of large flat areas that lack specific ligand binding pockets, making them challenging binding sites for small molecules [6]. Moreover, while PPI hotspot detection methods offer valuable insights into PPIs, they will overlook other important areas of the protein. In the pursuit of accurately identifying the most critical regions of proteins, we investigated a new approach based on recently developed artificial intelligence (AI) models: OmegaFold (OF) for predicting protein folding and AlphaMissense (AM) for predicting the pathogenicity of missense variants.

Recently, several studies have emphasized the potential of AM for the detection of functionally important regions within proteins [7, 8]. AlphaMissense is an AI model derived from the AlphaFold protein folding algorithm (AF). It uses information on the structural context and evolutionary conservation to forecast the pathogenicity of missense variants. AM does not predict the structural changes of the mutated amino acid sequences but instead predicts pathogenicity as scalar values, ranging from 0 to 1, reflecting the likelihood of a specific missense mutation being pathogenic [9].

Besides, recent research has highlighted the potential of protein language models and protein folding models in offering valuable information, either through embeddings or through subsequent structural feature extraction, for predicting PPI hotspots [3] as well as variant pathogenicity [10]. Although AF is regarded as more accurate than OF for the prediction of protein structure, some reports have suggested that it is unable to accurately predict the structural changes upon point mutation [11]. Moreover, OF demonstrates quicker inference time than AF [12], which is advantageous when the inference of a large number of models is required.

Received: September 13, 2024. **Revised:** March 26, 2025. **Editorial Decision:** April 9, 2025. **Accepted:** April 10, 2025

We developed a novel method that integrates both AM and OF to predict protein structural hotspots and evaluated the accuracy of this approach on the DAX-1 protein, which has been extensively studied regarding the functional effect of its variants [13] and has been associated with human diseases such as X-linked adrenal hypoplasia congenita (AHC) and hypogonadotropic hypogonadism [14, 15]. Remarkably, all DAX-1 missense AHC mutations studied to date have in common the effect of shifting the localization of the protein to the cytoplasm, thereby impairing the transcriptional repression activity of the wild-type (WT) protein [13, 16].

Materials and methods

With the aim of improving the explainability of AM predictions, we used a distillation approach involving an XGBoost regression model. As a tree-based method, XGBoost allows straightforward access to explainability through feature importance, which can be summed for each position in the protein sequence to identify the most critical residues (Fig. 1). Briefly, the structure of the single amino acid variants of the protein is inferred using OF [12], and the structures of variants are analyzed using the local distance difference test (lDDT) [17] and dictionary of secondary structure in proteins (DSSP) [18] to yield structural features, which are in turn used to train the XGBoost regressor model to predict the AM pathogenicity score of each variant [9]. The XGB model hyperparameters are optimized automatically with the Optuna library, and the feature importance of the trained model is summed and normalized for each residue position, to yield a new metric we will refer to as the "AMX score," which allows the detection of the most important structural regions of a protein.

Software and resources

The following software versions were used: Python 3.10, xgboost 2.1.0, lddt 2.2, and dssp 4.2.2. Mol* Viewer was used for the visualization of three-dimensional (3D) models [19].

Computation of the 3D structures for the 8930 single amino acid variants of DAX-1 with OF required ∼9 days on one GPU (NVIDIA RTX4090), and the rest of the analysis pipeline required ∼1 h on CPU (AMD Ryzen Threadripper PRO 5955WX 16-Cores).

Datasets

Dataset of clinical significances of DAX-1 variants for the assessment of AM performance

The clinical significance of DAX-1 variants was retrieved from the European Bioinformatics Institute website, which aggregates data from UniProt, ClinVar, and Ensembl (https://www.ebi.ac.uk/proteins/api/variation/P51843). Among 533 entries in the database, 86 were single amino acid variants with a reported clinical significance. Entries with clinical significance of types "Pathogenic" and "Likely pathogenic," as well as "Benign" and "Likely benign," were pooled together, respectively, yielding only three final classes: "Benign" ($n = 21$), "Ambiguous" ($n = 24$), and "Pathogenic" ($n = 41$).

Dataset for AM distillation with XGBoost

Inference of protein variant structures

As the new AMX approach proposed in the present work requires the prediction of many protein structures (8930 single amino acid variants for the 470 residues long DAX-1, accounting for 19 possible substitutions per position within the 20 standard amino acids), we used OF for the folding prediction, owing to its relatively lightweight computing requirements and good accuracy [12]. Notably, predictions made by OF and AlphaFold2 for the structured part of the human DAX-1 protein exhibit a high degree of similarity (Supplementary Fig. S1), as evidenced by the values of the TM score and lDDT similarity metrics, which are very close to 1 when comparing these models (0.97 and 0.92, respectively). All the protein structures originating from *in silico* predictions used throughout this study were computed with OF. The canonical sequence of DAX-1 was retrieved from UniProt (ID: P51843-1), and the structures of each of its single amino acid variants, obtained by substitution with one of the 20 standard amino acids, were inferred with OF using default parameters.

Feature extraction from variant structures

The lDDT was obtained by analyzing structures of variants with lddt 2.2, to obtain the lDDT score for each position in the sequence, using the structure predicted for the WT sequence as a reference, and default lddt parameters: all pairs of atoms within a 15 Å inclusion radius and not belonging to the same residues were considered [17]. The predicted lDDT (plDDT) was retrieved for each variant, from the *B* factor field of the pdb structure files. The plDDT is the prediction, made by the folding model itself, of the lDDT-Cα, which only considers distances between carbon α atoms. It ranges from 0 to 100, with values <50 usually indicating disordered regions [20]. Backbone angles Phi and Psi and the solvent accessibility area per residue (RSA) were retrieved using dssp 4.4.2 [18]. The feature values obtained for each residue position were concatenated to make a single vector for each variant, serving as input to the XGB regressor model.

Other features, such as flexibility [21], hydrophobicity [22], amino acid type, and hydrogen bond parameters derived from DSSP, were also evaluated before finally selecting the set of the five most useful features: lDDT, plDDT, Psi, Phi, and RSA (Fig. 2A). The predictive performance of the reduced set of five features was comparable to that of the full feature set, as indicated by the root mean squared error (RMSE) for AM score predictions, with respective RMSE values of 0.179 ± 0.010 and 0.173 ± 0.012 (Fig. 2B).

Retrieving AlphaMissense score of variants

The pathogenicity scores predicted by AM for each variant were used as targets for training of the XGB regressor, and were retrieved from the catalogue of all human single amino acid substitutions, available at https://console.cloud.google.com/storage/browser/dm_alphamissense. The AM scores are single values ranging from 0 to 1 and can be interpreted as the approximate probability of a variant being clinically pathogenic [9]. AlphaMissense distinguishes three pathogenicity classes based on the AM score: likely benign (<0.34), likely pathogenic (>0.564), and ambiguous ($0.34 \leq$ AM score ≤ 0.564).

XGBoost regression model training and hyperparameter tuning

The XGB regression model was trained using the open-source XGBoost library [23]. This kind of model belongs to the family of decision tree ensembles. The principle of boosting at the core of XGBoost models (eXtreme Gradient Boosting) consists in building trees sequentially one after another during

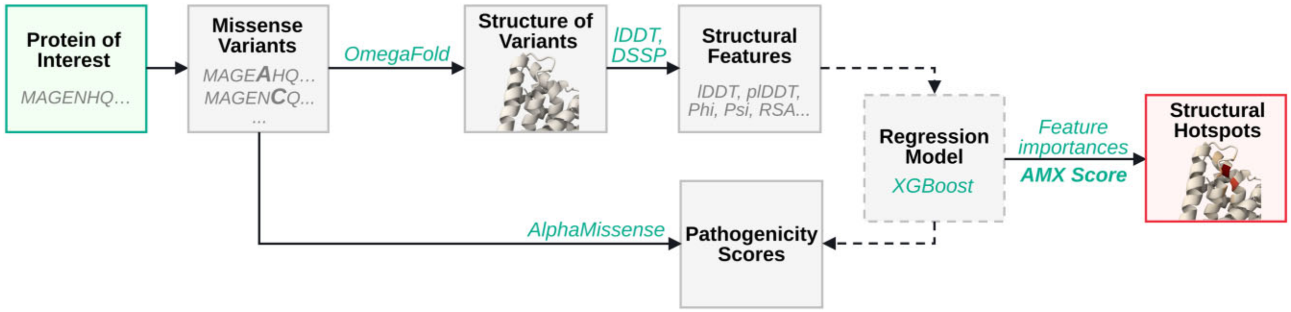

Figure 1. Overview of the method for detecting structural hotspots through the calculation of the AMX score of a protein.

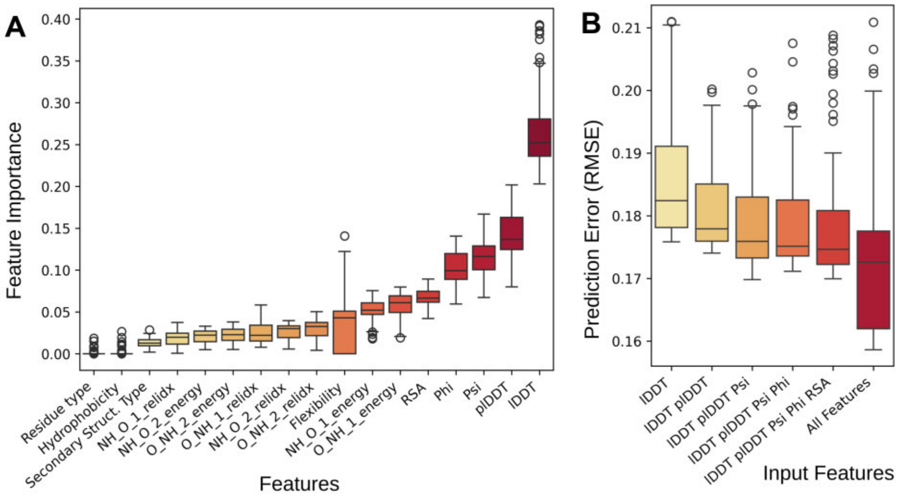

Figure 2. (**A**) Importance of different features, observed on 100 XGBoost regressors trained during the hyperparameter tuning. (**B**) RMSE on the prediction of AM score by XGB regressors trained with different sets of input features.

fitting, with each new tree aiming to correct the errors of the previous ones.

The hyperparameters of the XGB model were tuned using the Optuna framework [24]. The range of values explored for each parameter is given in Supplementary Table S1. One hundred rounds of evaluations were performed, to identify the set of hyperparameters minimizing the RMSE on the prediction of the AM scores of variants.

Extraction of feature importances and AMX score

Models based on decision trees can provide importance scores for each input feature, indicating how useful a feature was in constructing the model's trees. We used the feature importance based on gain, as defined here: https://xgboost.readthedocs.io/en/latest/tutorials/model.html. The AMX score of a given position is then obtained as the normalized sum of all feature importance for this position.

Results

Evaluation of AlphaMissense performances on DAX-1

Before the evaluation of the AMX approach for the detection of functional hotspots, we examined the predictions of AlphaMissense on the DAX-1 protein. The heatmap generated by plotting the AM scores for different amino acid substitutions along the sequence of DAX-1 (Fig. 3A) reveals distinct regions characterized by elevated pathogenic risk associated with single amino acid variants. Specifically, within the ligand binding domain (LBD) spanning positions 205–470, the AM pathogenicity scores consistently exhibit high values, with the exception of three disordered regions: the N-terminal segment of the LBD (residues 205–245), the H5–H7 loop (residues 311–352), and the H9–H10 loop (residues 418–424). In the N-terminal domain (NTD), occupying positions 1–204, the regions showing elevated AM pathogenicity scores are scarcer and primarily located around two LXXML motifs (residues 13–17 and 80–84), one LXXLL motif (residues 146–150), and four cysteine-rich regions (around positions 38–43, 64–69, 104–110, and 130–145). To gain further insights, the 3D structure of DAX-1 was visualized with colorization based on the average AM pathogenicity scores per position (Fig. 3B). Notably, positions with higher AM scores tend to cluster within the core of the LBD, frequently involving residues that face the center of the core or interact with other alpha-helices.

We also assessed the precision of AM predictions on a curated dataset of DAX-1 variants with known clinical significances, extracted from the EMBL-EBI database (Supplementary Table S2). AM exhibited 78% accuracy (32/41) for predicting pathogenic variants and 100% accuracy (21/21) for predicting benign variants, based on our curated EMBL-EBI dataset of DAX-1 variants (Fig. 3C). In contrast, the ambiguous variants were correctly identified with only 12.5% accuracy (3/24). Among these variants, 75% (18/24) were classified as benign. To comprehen-

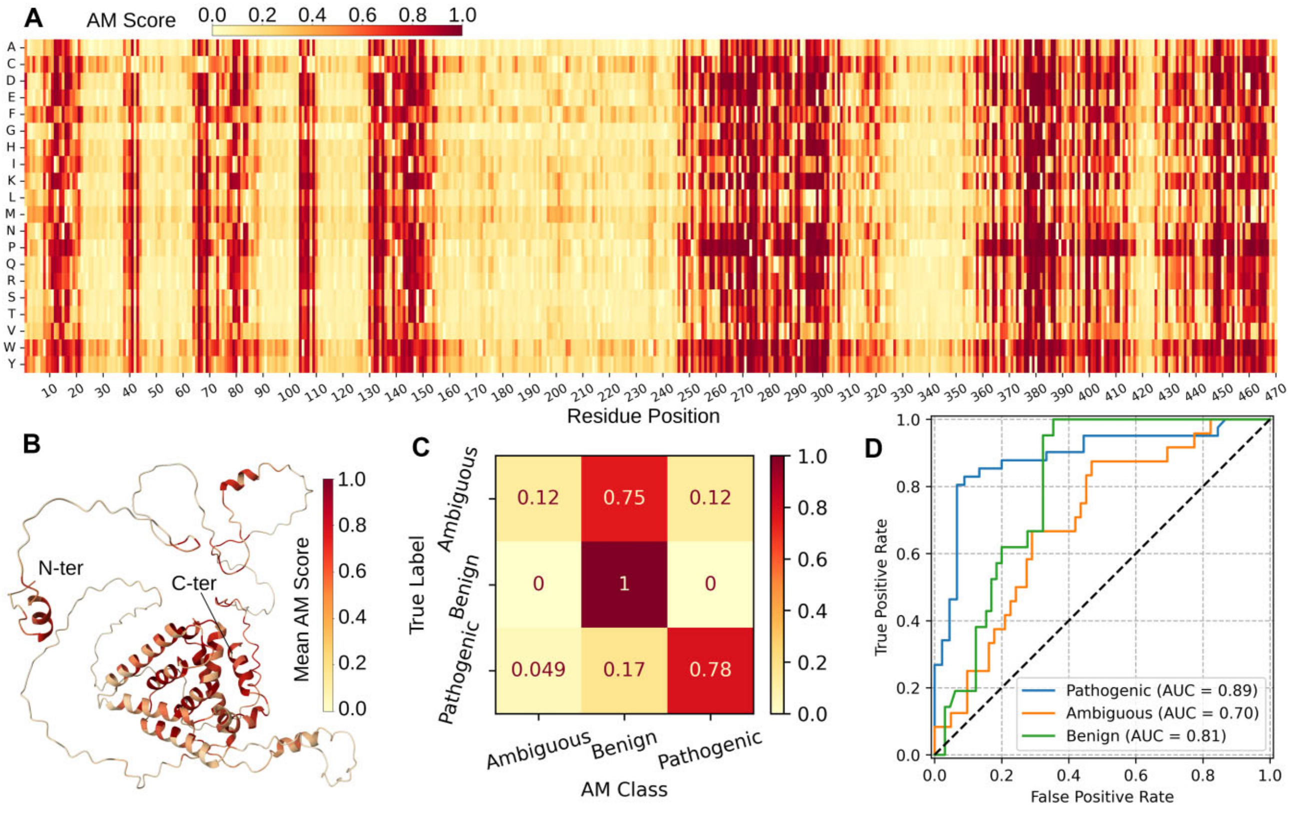

Figure 3. (**A**) Pathogenicity heatmap obtained with AM scores of all single amino acid variants of the human DAX-1 protein. (**B**) OF predicted structure of human DAX-1, with positions colored by their mean AM score. (**C**) Confusion matrix and (**D**) receiver operating characteristic (ROC) curve of AM predictions benchmarked against 86 single amino acid variants of DAX-1 from the EMBL-EBI database; individual ROC for each pathogenicity class was obtained by considering the relevant class as the true positive class.

sively evaluate the predictive performance of AM, we constructed an ROC curve considering all pairwise comparisons of pathogenicity predictions made by AM (Fig. 3D). In this evaluation, each one of the possible outcomes (pathogenic, ambiguous, or benign) was considered as a true positive, while treating the other two predictions as false positives. The pathogenic predictions presented the highest area under the curve (AUC) of 0.89, followed by benign predictions with an AUC of 0.81, and ambiguous predictions with an AUC of 0.70.

Comparison of hotspot detection with AM and AMX, in the LBD of DAX-1

The inspection of the mean AM score in the LBD (Fig. 4A) reveals the presence of three clusters with elevated pathogenicity risks, broadly corresponding with the structured core of the LBD. On a finer level, the repressor helix (RH, 273–279) identified in the mouse Dax-1 LBD [25] and RH-facing residues of H9 (Tyr399, Ile400, and Leu403) both exhibit high mean AM scores (Fig. 4C); however, these areas are overshadowed by other larger clusters of high-scoring positions, located in helices H8 (residues 376–383), H5 (residues 294–302), and H12 (residues 458–467).

In contrast to the widespread dispersion of positions with a high mean AM score, the distribution of positions with a high AMX score appears to be significantly more concentrated, with the most prominent positions found on the H4 helix (Gln282, Leu286, and Cys290), the C-terminal part of H8 (Val385), and the N-terminal part of the H8–H9 loop (Asn388, Val391), as illustrated in Fig. 4B. Additionally, despite being seemingly scattered along the sequence, all these positions are spatially clustered within a single region, which exhibits close interaction with the RH and N-terminal segment of H9 (Fig. 4D).

In the N-terminal domain of DAX-1 (1–204), the AMX scores are generally low except in the vicinity of the first LXXML motif (residues 13–17), with the highest AMX score for the residues Asn15 and Ser19 (Fig. 5). Also interestingly, the main contributing feature to elevated AMX scores observed in the LBD is the lDDT, whereas it is the plDDT in the case of the NTD (Supplementary Fig. S2). The lDDT reflects the local similarity between the variant structure and the WT structure, whereas the plDDT can be seen as an estimate of the local degree of order or of flexibility. Thus, a high importance of lDDT implies that preservation of local similarity to the WT structure is critical for activity, while a high importance of plDDT implies preservation of local flexibility is critical. These two properties are slightly different, as the WT structure itself may be preserved in a given region of a variant, while at the same time, local flexibility may be lost.

Experimental evidence for the functional relevance of the hotspot identified with AMX

We mapped the mean AM and AMX scores to an experimentally determined structure of the mouse Dax-1 LBD bound to one of its physiological targets (the liver receptor homolog 1 nuclear receptor [LRH-1, RCSB structure 3F5C]), taking into account a shift between the human and mouse sequences

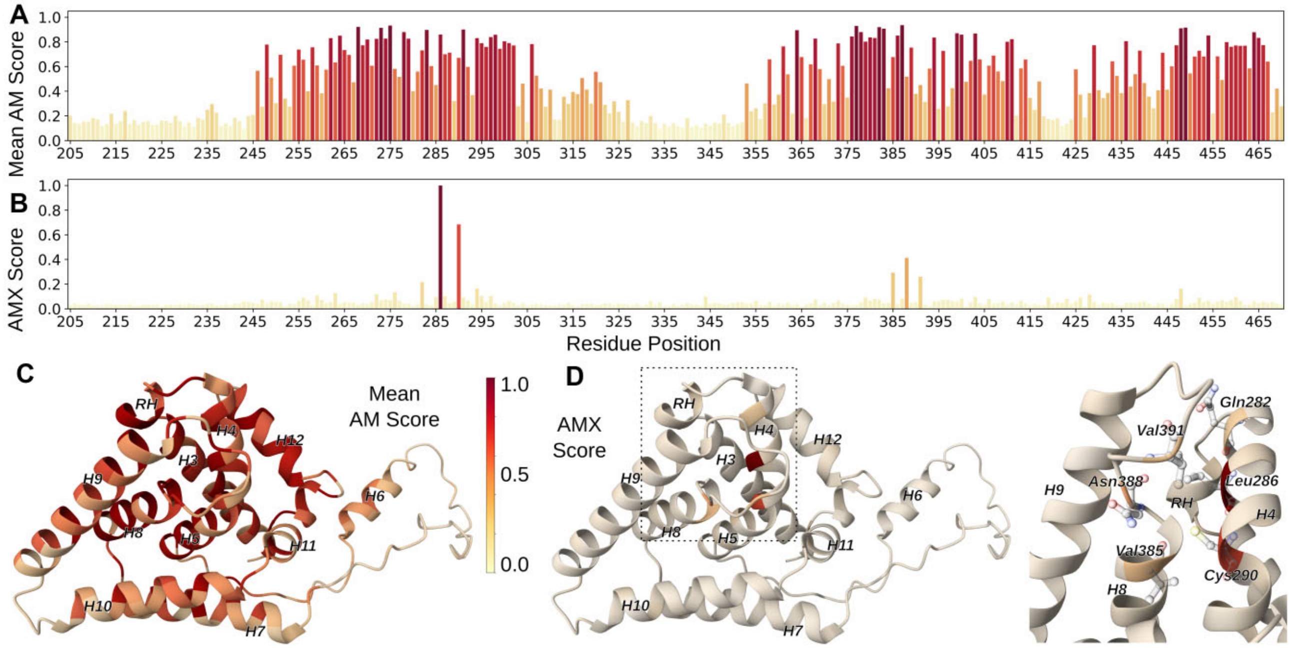

Figure 4. (**A**) Mean AM score and (**B**) AMX score, by residue position, in the human DAX-1 LBD. (**C**) OF predicted structure of human DAX-1, with positions colored by their mean AM score (shown residues: 250–470). (**D**) OF model of human DAX-1, with residue positions colored by their AMX importance score, and close-up view of the hotspot revealed by AMX score.

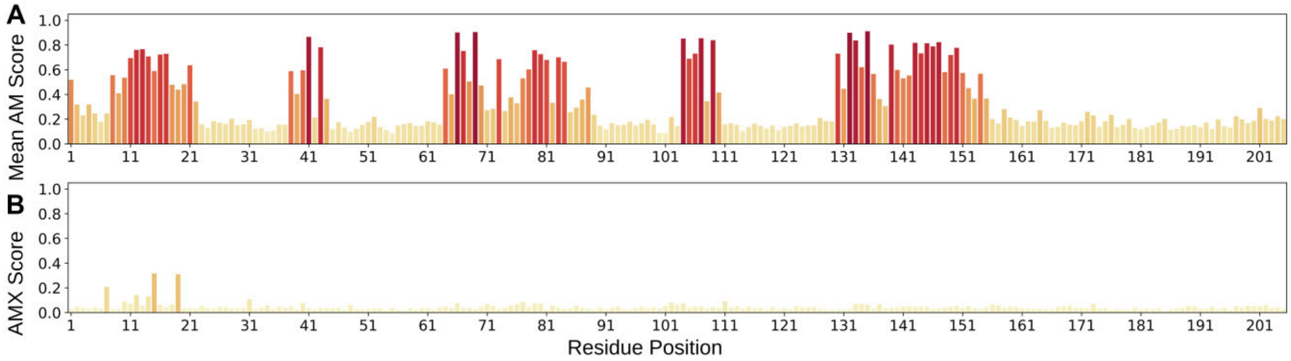

Figure 5. (**A**) Mean AM score and (**B**) AMX score, by residue position, in the NTD of DAX-1.

caused by the insertion of two alanine residues in the NTD of the murine DAX-1 at positions 103 and 113 (UniProt ID: Q61066, alignment with P51843 in Supplementary Fig. S3). As can be seen in Fig. 4, there does not appear to be a clearly defined hotspot when inspecting the DAX-1 model colorized with mean AM score (Fig. 6A), whereas the model colorized with the AMX score highlights a single region centered around the leucine 288 residue (Leu286 in human DAX-1, Fig. 6B). This hydrophobic residue is in close interaction with other hydrophobic residues of the RH and N-terminal part of H9, thus likely playing an important role in the stabilization of the PPI with LRH-1, which is known to involve these two helices [25].

Discussion

By analyzing the feature importance of an XGB model trained to predict AM pathogenicity score based on structural features of variants, the AMX score provides a clear view of the most important regions of DAX-1 structure for the conservation of its function. This claim is supported by the fact that the positions highlighted by the AMX score are concentrated in a few spatial regions that are known to be important for the function of DAX-1.

Indeed, the hotspot identified in DAX-1 LBD, located between the H4 and the H8–H9 loop, underpins its PPI with LRH-1 [25], making this region crucial for the integrity of DAX-1 functions. Interestingly, the main contributing feature for the elevated AMX score in this region is the lDDT, which is obtained by comparing the local environment of a residue in between the wild type versus missense variant structures. As the lDDT is computed for each atom of a residue with a 15 Å inclusion radius for considering interatomic distances, it implies that preservation of the native structure of this region and its immediate surroundings is important for the function of DAX-1, which is in agreement with the observation that this region supports and stabilizes the interface between DAX-1 and LRH-1.

Unsurprisingly, this region also harbors established disease-causing mutations associated with AHC, such as Leu278Pro

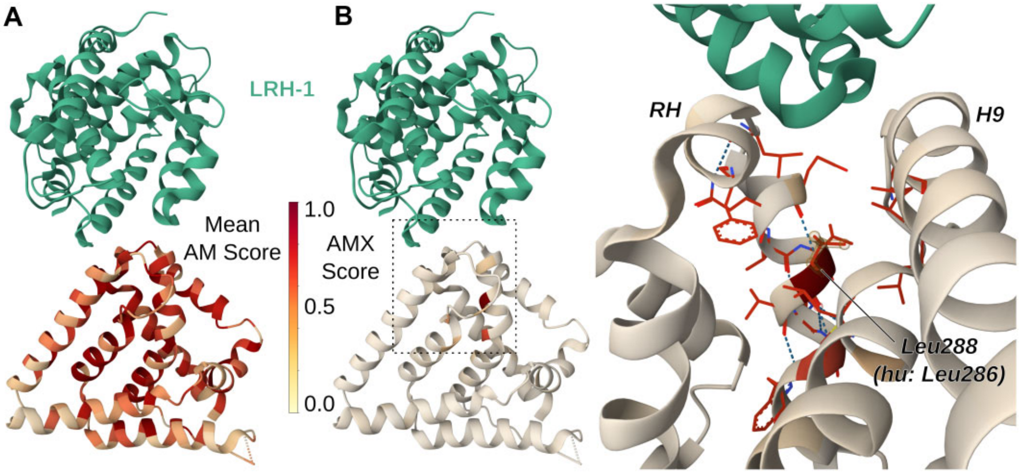

Figure 6. Murine LRH-1 (green) in complex with murine DAX-1 (chain C of RCSB model 3F5C). (**A**) DAX-1 colorized with mean AM score per residue. (**B**) DAX-1 colorized with the AMX scores (computed from human DAX-1 variants) and close-up view of the surroundings of L288 (corresponding to L286 in human DAX-1).

[26] in RH and Val287Gly [27] in H4. These residues have in common the orientation of their side chain, both pointing toward the hotspot identified using AMX. In contrast, mutations affecting residues of H4 that are facing away from this structural hotspot, such as Leu280Arg, Leu284Arg, or Arg288Ala, have been shown to have negligible impact on the nuclear localization or transcriptional repression activity of DAX-1 [16].

Additionally, the Val385Gly mutation, which impacts one of the highest-scoring positions identified by AMX in our study, is associated with AHC [28] and has been demonstrated to cause a drastic loss of nuclear localization and physiological activity of DAX-1 [13]. The abnormal localization of this variant despite the presence of intact nuclear localization signals might be explained by the misfolding and aggregation of the mutated protein in the cytosol [16], further supporting the structural importance of the hotspot detected with AMX.

A second hotspot presenting an elevated AMX score, found in the NTD of DAX-1, overlaps with the first LXXML motif (residues 13–17). This sequence is an NR-box-like motif, a variant of the NR-box, whose consensus sequence is LXXLL [29]. The NR-box enables specific interaction with the AF-2 domain of nuclear receptors by adopting an alpha helix conformation upon binding [30]. The ability of these motifs to undergo disorder-to-order transition is thought to be key for their regulatory activity, allowing them to form transient, reversible complexes [31]. This aligns with our findings in DAX-1, where the main contributing feature to the heightened AMX score of this motif is the plDDT, which is known to correlate with the local degree of order [32], underscoring the importance of this motif's intrinsic structural order for DAX-1 function.

Our method mainly relies on structural features, which have demonstrated greater effectiveness in predicting the AlphaMissense pathogenicity score for DAX-1 compared to sequence-based features such as residue type, hydrophobicity, or flexibility (Fig. 2A). We posit that this enhanced performance is because structural features can incorporate information from protein folding predictions, an aspect absent from purely sequence-based features. Moreover, these structural features (lDDT, plDDT, Phi, Psi, and RSA) are capturing local structural characteristics, making them relatively insensitive to large-scale structural changes and disorder. Nevertheless, future research focusing on intrinsically disordered proteins will be necessary to assess the ability of the AMX approach to detect hotspots in unstructured regions of proteins.

The main limitation of this approach is the substantial computational resources needed to predict the 3D structures of variants. Larger-scale applications studying multiple proteins, or longer proteins, may require high-performance computing resources with the current approach. To mitigate this, OF could be replaced by faster folding models like ESMFold, or only a subset of variants could be considered. An informed strategy for the subsampling of variants could be based on a reduced set of amino acids, akin to the approaches used in mutagenesis studies [33], aiming to minimize the number of representative mutations that require prediction. These strategies would require further study, as they might improve speed, but could compromise accuracy. Meanwhile, our results demonstrate the feasibility of this approach with consumer-grade hardware (NVIDIA RTX4090 GPU), even for a medium-sized protein like DAX-1.

To study the generalizability of the AMX approach, we also applied it to the p53 protein (Supplementary Fig. S4). Similar to the results observed for DAX-1, the AMX score identified a distinct hotspot located in the protein core, while the AM scores were generally elevated across all structured regions of the protein. For p53, this hotspot comprised residues 137–139 and 235–236, which are involved in the ubiquitination of p53 and its DNA binding activity [34]. Moreover, some of these residues, such as Lys139, are highly conserved [35]. The significance of the detected region for p53's function further suggests the generalizability of the AMX approach. However, we reckon that our method would need larger-scale benchmarking to confirm its general applicability.

Our objective was to identify the most structurally important areas of a protein, by leveraging the explainability of an XGBoost surrogate model of AlphaMissense, trained on structural features of protein variants. While the AM score effectively predicts the pathogenicity of single amino acid variants in DAX-1, it lacks precision in pinpointing structural hotspots, and the analysis of AM scores highlights ordered regions of the protein without a specific focus on particular regions. On the other hand, the AMX score provides a much clearer depiction of the most important protein domains. The differences in AMX score across the protein structure reveal spatially defined regions of high-scoring residues, whose structural characteristics were confirmed to be significant for protein function.

The AMX score is a relative measurement, aimed at prioritizing regions crucial for protein function rather than acting as an absolute indicator of residue essentiality for protein function. Our analysis on DAX-1 effectively identifies primary hotspots, while also revealing secondary hotspots with slightly elevated values. Defining a threshold value, below which the prioritization becomes less significant, represents an important objective for future research. Studies incorporating larger datasets of proteins are expected to offer the necessary guidelines for a more nuanced interpretation of the AMX score. Nevertheless, we remain confident in the current applicability of the AMX approach for identifying the most critical regions underpinning a protein's function. This knowledge could offer direction for the design of experiments probing the function of proteins and for the development of pharmaceuticals intended to modulate protein function.

Acknowledgements

Author contributions: Noé Dumas (Conceptualization [equal], Investigation [equal], Methodology [equal], Software [equal], Validation [equal], Visualization [equal], Writing—original draft [equal], Writing—review & editing [equal]), Geoffrey Portelli (Conceptualization [equal], Methodology [equal], Resources [equal], Software [equal], Supervision [equal], Writing—original draft [equal], Writing—review & editing [equal]), Yang Ji (Formal analysis [equal], Methodology [equal], Software [equal]), Florent Dupont (Conceptualization [equal], Funding acquisition [equal], Methodology [equal], Project administration [equal], Supervision [equal], Writing—original draft [equal], Writing—review & editing [equal]), Mehdi Jendoubi (Conceptualization [equal], Funding acquisition [equal], Project administration [equal]), and Enzo Lalli (Conceptualization [equal], Supervision [equal], Validation [equal], Writing—original draft [equal], Writing—review & editing [equal])

Supplementary data

Supplementary data is available at NAR Genomics & Bioinformatics online.

Conflict of interest

None declared.

Funding

Agence Nationale de la Recherche (ANR) grant ANR-20-CE14-0007, awarded to E.L; Thales internal R&D funds for innovative projects.

Data availability

The code and data underlying this article are available on Zenodo at https://zenodo.org/records/14860715. An executable version of the code can be found online under the form of a Google Colab notebook at https://colab.research.google.com/drive/100×2rLfIhmwOvfQbe-es38LxPtFJWOBp?usp=sharing. This notebook has the ability to perform AMX analysis on any protein listed in the AM catalogue, requiring only the structural data of the WT protein and its variants in PDB format. Additionally, the Zenodo repository offers supplementary datasets, beyond the examples of DAX-1 and p53, that are ready for analysis with the Colab notebook.

References

1. Glaser F, Pupko T, Paz I *et al.* ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. *Bioinformatics* 2003;**19**:163–4. https://doi.org/10.1093/bioinformatics/19.1.163
2. Chen J, Kuhn LA, Raschka S. Techniques for developing reliable machine learning classifiers applied to understanding and predicting protein:protein interaction hot spots. In: Gore M, Jagtap UB (eds.), *Computational Drug Discovery and Design*, New York, NY: Springer US, 2024, 235–68.
3. Sargsyan K, Lim C. Using protein language models for protein interaction hot spot prediction with limited data. *BMC Bioinformatics* 2024;**25**:115. https://doi.org/10.1186/s12859-024-05737-2
4. Moreira IS, Koukos PI, Melo R *et al.* SpotOn: high accuracy identification of protein–protein interface hot-spots. *Sci Rep* 2017;**7**:8007. https://doi.org/10.1038/s41598-017-08321-2
5. Wang H, Liu C, Deng L. Enhanced prediction of hot spots at protein–protein interfaces using extreme gradient boosting. *Sci Rep* 2018;**8**:14285. https://doi.org/10.1038/s41598-018-32511-1
6. Xie X, Yu T, Li X *et al.* Recent advances in targeting the "undruggable" proteins: from drug discovery to clinical trials. *Signal Transduct Target Ther* 2023;**8**:335. https://doi.org/10.1038/s41392-023-01589-z
7. Tordai H, Torres O, Csepi M *et al.* Analysis of AlphaMissense data in different protein groups and structural context. *Sci Data* 2024;**11**:495. https://doi.org/10.1038/s41597-024-03327-8
8. McDonald EF, Oliver KE, Schlebach JP *et al.* Benchmarking AlphaMissense pathogenicity predictions against cystic fibrosis variants. *PLoS One* 2024;**19**:e0297560. https://doi.org/10.1371/journal.pone.0297560
9. Cheng J, Novati G, Pan J *et al.* Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science* 2023;**381**:eadg7492. https://doi.org/10.1126/science.adg7492
10. Schmidt A, Röner S, Mai K *et al.* Predicting the pathogenicity of missense variants using features derived from AlphaFold2. *Bioinformatics* 2023;**39**:btad280. https://doi.org/10.1093/bioinformatics/btad280
11. Buel GR, Walters KJ. Can AlphaFold2 predict the impact of missense mutations on structure? *Nat Struct Mol Biol* 2022;**29**:1–2. https://doi.org/10.1038/s41594-021-00714-2
12. Wu R, Ding F, Wang R *et al.* High-resolution *de novo* structure prediction from primary sequence. bioRxiv, https://doi.org/10.1101/2022.07.21.500999, 22 July 2022, preprint: not peer reviewed.
13. Lehmann SG, Lalli E, Sassone-Corsi P. X-linked adrenal hypoplasia congenita is caused by abnormal nuclear localization of the DAX-1 protein. *Proc Natl Acad Sci USA* 2002;**99**:8225–30. https://doi.org/10.1073/pnas.122044099
14. Lalli E, Bardoni B, Zazopoulos E *et al.* A transcriptional silencing domain in DAX-1 whose mutation causes adrenal hypoplasia

congenita. *Mol Endocrinol* 1997;**11**:1950–60. https://doi.org/10.1210/mend.11.13.0038

15. Lalli E. Role of orphan nuclear receptor DAX-1/NR0B1 in development, physiology, and disease. *Adv Biol* 2014;**2014**:1–19. https://doi.org/10.1155/2014/582749
16. Lehmann SG, Wurtz J-M, Renaud J-P *et al.* Structure–function analysis reveals the molecular determinants of the impaired biological function of DAX-1 mutants in AHC patients. *Hum Mol Genet* 2003;**12**:1063–72. https://doi.org/10.1093/hmg/ddg108
17. Mariani V, Biasini M, Barbato A *et al.* lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics* 2013;**29**:2722–8. https://doi.org/10.1093/bioinformatics/btt473
18. Joosten RP, Beek TAH, Krieger E *et al.* A series of PDB related databases for everyday needs. *Nucleic Acids Res* 2011;**39**:D411–9. https://doi.org/10.1093/nar/gkq1105
19. Sehnal D, Bittrich S, Deshpande M *et al.* Mol* Viewer: modern web app for 3D visualization and analysis of large biomolecular structures. *Nucleic Acids Res* 2021;**49**:W431–7. https://doi.org/10.1093/nar/gkab314
20. Guo H-B, Perminov A, Bekele S *et al.* AlphaFold2 models indicate that protein sequence determines both structure and dynamics. *Sci Rep* 2022;**12**:10696. https://doi.org/10.1038/s41598-022-14382-9
21. Vihinen M, Torkkila E, Riikonen P. Accuracy of protein flexibility predictions. *Proteins* 1994;**19**:141–9. https://doi.org/10.1002/prot.340190207
22. Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. *J Mol Biol* 1982;**157**:105–32. https://doi.org/10.1016/0022-2836(82)90515-0
23. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY: Association for Computing Machinery, 2016, 785–94.
24. Akiba T, Sano S, Yanase T *et al.* Optuna: a next-generation hyperparameter optimization framework. In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD '19*. New York, NY: Association for Computing Machinery, 2019, 2623–31.
25. Sablin EP, Woods A, Krylova IN *et al.* The structure of corepressor Dax-1 bound to its target nuclear receptor LRH-1. *Proc Natl Acad Sci USA* 2008;**105**:18390–5. https://doi.org/10.1073/pnas.0808936105
26. Bassett JH, O'Halloran DJ, Williams GR *et al.* Novel DAX1 mutations in X-linked adrenal hypoplasia congenita and hypogonadotrophic hypogonadism. *Clin Endocrinol* 1999;**50**:69–75. https://doi.org/10.1046/j.1365-2265.1999.00601.x
27. Franzese A, Brunetti-Pierri N, Spagnuolo MI *et al.* Inappropriate tall stature and renal ectopy in a male patient with X-linked congenital adrenal hypoplasia due to a novel missense mutation in the DAX-1 gene. *Am J Med Genet A* 2005;**135**:72–4. https://doi.org/10.1002/ajmg.a.30670
28. Zhang YH, Guo W, Wagner RL *et al.* DAX1 mutations map to putative structural domains in a deduced three-dimensional model. *Am Hum Genet* 1998;**62**:855–64. https://doi.org/10.1086/301782
29. Heery DM, Kalkhoven E, Hoare S *et al.* A signature motif in transcriptional co-activators mediates binding to nuclear receptors. *Nature* 1997;**387**:733–6. https://doi.org/10.1038/42750
30. Wärnmark A, Treuter E, Wright APH *et al.* Activation functions 1 and 2 of nuclear receptors: molecular strategies for transcriptional activation. *Mol Endocrinol* 2003;**17**:1901–9. https://doi.org/10.1210/me.2002-0384
31. Fuxreiter M, Tompa P, Simon I. Local structural disorder imparts plasticity on linear motifs. *Bioinformatics* 2007;**23**:950–6. https://doi.org/10.1093/bioinformatics/btm035
32. Tunyasuvunakool K, Adler J, Wu Z *et al.* Highly accurate protein structure prediction for the human proteome. *Nature* 2021;**596**:590–6. https://doi.org/10.1038/s41586-021-03828-1
33. Chen MMY, Snow CD, Vizcarra CL *et al.* Comparison of random mutagenesis and semi-rational designed libraries for improved cytochrome P450 BM3-catalyzed hydroxylation of small alkanes. *Protein Eng Des Sel* 2012;**25**:171–8. https://doi.org/10.1093/protein/gzs004
34. Rainwater R, Parks D, Anderson ME *et al.* Role of cysteine residues in regulation of p53 function. *Mol Cell Biol* 1995;**15**:3892–903. https://doi.org/10.1128/MCB.15.7.3892
35. Chan WM, Mak MC, Fung TK *et al.* Ubiquitination of p53 at multiple sites in the DNA-binding domain. *Mol Cancer Res* 2006;**4**:15–25. https://doi.org/10.1158/1541-7786.MCR-05-0097

Received: September 13, 2024. **Revised:** March 26, 2025. **Editorial Decision:** April 9, 2025. **Accepted:** April 10, 2025