PLOS COMPLEX SYSTEMS

RESEARCH ARTICLE

A single-graph visualization to reveal hidden explainability patterns of SHAP feature interactions in machine learning for biomedical issues

Félix Furger[1☯], Miguel Thomas[1☯], Julien Aligon[2], Emmanuel Doumard[1,2], Cyrille Delpierre[3], Louis Casteilla[1], Paul Monsarrat[1,4¤*]

1 RESTORE Research Center, Universite de Toulouse, INSERM 1301, CNRS 5070, EFS, Toulouse, France, **2** Universite Toulouse Capitole, Institute of Research in Informatics (IRIT) of Toulouse, CNRS – UMR5505, Toulouse, France, **3** CERPOP, UMR1295 (EQUITY), Universite de Toulouse, Toulouse, France, **4** Oral Medicine Department and Hospital of Toulouse, Toulouse, France

☯ These authors contributed equally to this work.
¤ Current address: RESTORE Research Center, 4bis Avenue Hubert Curien 31100 Toulouse, France
* paul.monsarrat@inserm.fr

OPEN ACCESS

Citation: Furger F, Thomas M, Aligon J, Doumard E, Delpierre C, Casteilla L, et al. (2025) A single-graph visualization to reveal hidden explainability patterns of SHAP feature interactions in machine learning for biomedical issues. PLOS Complex Syst 2(9): e0000060. https://doi.org/10.1371/journal.pcsy.0000060

Editor: Juan Gonzalo Barajas-Ramirez, IPICYT: Instituto Potosino de Investigacion Cientifica y Tecnologica AC, MEXICO

Received: January 08, 2025

Accepted: July 14, 2025

Published: September 4, 2025

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pcsy.0000060

Data availability statement: More details about data availability for Physiological Age can be

Abstract

In the last decades, the utility of Machine Learning (ML) in the biomedical domain has been demonstrated repeatedly. Their inherent opacity need augmenting ML with explainability techniques. A common practice in model explainability however, is to focus solely on the explanatory values themselves without accounting for both the main and interaction effects. While this approach simplifies interpretation, it potentially overlooks critical medical information since the nature of the interactions may provide clues to the underlying biological mechanisms. This article introduces a novel method for analyzing explanatory values of machine learning (ML) models, in the form of a comprehensive graphical visualization. The method not only emphasises the individual contributions of the features but also gives insights about the interactions they share with one another. Designed for local additive explanation methods, the proposed tool effectively translates the complex and multidimensional nature of these values into an intuitive single-graph format. It offers a clear window into how feature interactions contribute to the overall prediction of the ML model while aiding in the identification of various interaction types, such as mutual attenuation, positive/negative synergies or dominance of one feature over another. This approach provides insights for generating hypotheses, improving the transparency of ML models, particularly in the context of biology and medicine since living organisms are characterised by a multitude of parameters in complex interactions, a complexity that ensures the “stability” and robustness of structures and functions.

found at Bernard D et al. (Aging Cell. 2023;22(8):e13872. doi:10.1111/acel.13872.). The SA-Heart dataset was also used Rossouw J et al. (South African Medical Journal. 1983;64(12):430–436) and downloaded from https://www.kaggle.com/datasets/waalbannyantudre/south-african-heart-disease-dataset.

Funding: This work was supported by fundings from Programme d'Investissements d'Avenir, the Agence Nationale pour la Recherche (grant EUR CARe N°ANR-18-EURE-0003) and the Agence Nationale pour la Recherche for the national infrastructure "ECELLFrance: Development of mesenchymal stem cell based therapies" (PIA-ANR-11-INBS-005). This work was also supported by Inserm Transfert.

Competing interests: The authors have declared that no competing interests exist.

Author summary

The use of machine learning (ML) represents a breakthrough in the biomedical field due to its ability to uncover complex relationships between features. SHAP explainability then makes it possible to understand the key factors that contributed to the prediction. This article introduces a novel method for analyzing explanatory values of ML models, in the form of a comprehensive graphical visualization. The method not only emphasizes the individual contributions of the features but also offers a more detailed view of how features interact with each other. This results in a directed graph where both interaction strength and directionality are encoded, enabling the discovery of higher-order patterns such as mutual attenuation or dominant influences. Such representations are particularly suited to biomedical systems, where understanding the interplay between variables is often key to interpreting complex physiological mechanisms.

Introduction

The need of explainability for the biomedical domain

In the last decades, the utility of Machine Learning (ML) in the biomedical domain has been demonstrated repeatedly [1,2]. Advanced models, including ensemble approaches and neural networks, have offered high predictive accuracy while leveraging increasingly complex and voluminous datasets. These advancements have significantly contributed to various breakthroughs in diagnostic accuracy and overall healthcare management [3]. However, using such sophisticated models often introduce one major drawback: their inherent opacity, often referred as the "black-box" dilemma [4]. This lack of transparency not only restricts the trust in and adoption of ML but also raises ethical concerns, particularly in sensitive sectors like healthcare, where decision justification for patients and practitioners is crucial [4,5]. In response to this challenge, emphasis has been placed on augmenting ML with explainability techniques in an attempt to shed light on the reasoning behind the algorithmic decisions [6]. This is all the more important as this explainability can also lead to a better understanding of pathophysiology and the underlying biological mechanisms, even if the inference of causality must be carefully considered.

Explainability techniques for ML

Explainability techniques can be broadly categorised into local and global methods. Local techniques focus on individual predictions, detailing why a model makes a specific decision, while global techniques provide an overview of the general model behavior. Interestingly, insights gained from local explanations can also be aggregated into global explanations to study both local and global behaviors of the model [6].

Examples of popular local explainability tools for ML are *LIME* [7] (Local Interpretable Model-Agnostic Explanations), *SHAP* [7] (SHapley Additive exPlanations), and coalition-based methods [6]. Such methods are described as being additive since for a single instance they produce a single vector of weights representing the contribution of each feature (including its interactions with other features) to the prediction. In each case, the contributions approximately sum up to the prediction minus the average prediction for the model. While each method has unique merits [6], *SHAP* stands out as one of the most utilised framework, notably for its ease of use, adaptability to diverse ML methods, and extensive library of functions and visualisation tools.

The intuition behind *SHAP* comes from cooperative game theory and the need for an equitable allocation of "payouts" among "players" using Shapley values. When transposing the idea to the context of ML, the "payouts" and the "players" take the form of the model predictions and the features, respectively. The exhaustive method, referred as the *complete method*, involves evaluating every possible coalition of features, with and without each feature. As coalitions are being evaluated, the impact on the prediction of the presence or absence of a feature in conjunction with other features is used to compute the feature's contribution [7]. Since the *complete method* is particularly expensive to compute, *SHAP* provides us with a more accessible solution creating perturbations to simulate the absence of a feature. *SHAP* unifies several XAI methods from the literature, in particular *LIME*, a linear local model used to approximate the change in the prediction. The resulting contributions exhibit the previously mentioned additive properties and are referred as SHAP values [7].

For a specific feature and a given sample, the SHAP value represents the additional contribution it provides to the prediction when combined with the set of features used to perform the prediction. By essence, for a given feature and a given prediction, the SHAP value hence encapsulates (i) the main effect, as the individual contribution of the feature when ignoring its interactions with other features and (ii) half the contributions arising from the interactions of the feature with all other features [7].

Interactions, often forgotten in the explanations

One of the major benefits of ML in the context of biology, and more particularly medicine, is its ability to approximate complex interactions between parameters, regardless of their nature. Indeed, living organisms are characterized by a multitude of parameters interacting in a complex way, a complexity that ensures the "stability" and robustness of structures and functions. A common practice in model explainability however, is to focus solely on the explanatory values themselves (*i.e.* SHAP values) without accounting for the decomposition between main and interaction effects. While this approach simplifies interpretation, it potentially overlooks critical medical information since the nature of the interactions (e.g. synergy, mutual attenuation) may provide clues to the underlying biological mechanisms that are not immediately apparent when examining individual predictors.

Despite the growing popularity of SHAP-based tools, current visualization methods such as summary plots, force plots, dependence plots, decision plots or interaction value plots remain limited when it comes to interpreting complex patterns of feature interactions. These visualizations either provide local insights that are difficult to aggregate, or global overviews that become unreadable as the dimensionality of the model increases. Several complementary tools—such as SHAPash [8], DALEX [9], InterpretML [10], or SHAP-IQ [11]—offer interactive dashboards or simplified feature-attribution views, but do not provide intuitive overviews of interaction dynamics. No existing method simultaneously visualizes both main effects and interaction effects of features in a unified, interpretable, and scalable format.

For a model with n features, there are $n(n–1)$ sets of SHAP interaction values as well as n sets of SHAP main effect values to analyze, resulting in a total of n^2 sets of data points to handle. Visualizing each of the n^2 SHAP dependence plots quickly becomes out of reach as n increases, hence complicating the task of finding patterns, groupings, and generalising findings for models with many features.

Consequently, this article introduces a novel graph-based visualization method that bridges this gap, offering a compact and biologically relevant representation of SHAP interaction patterns. The method not only emphasises the individual contributions of the features but also gives insights about the interactions they share with one another.

Materials and methods

Illustration context

In order to illustrate the proposed method, the explainable machine learning framework and database to predict physiological aging was considered [12]. This database involves 48 routine laboratory biological variables obtained for 60,322 individuals from the National Health and Nutrition Examination Survey (NHANES) database (1999 - 2018). All details were provided in the dedicated article [12]. As detailed previously, an *XGBoost* model was trained to predict chronological age based on the 48 biological variables as predictors (performance of 0.72 and 8.1 on the test dataset for R2 and MAE, respectively). The SHAP interactions values were subsequently extracted using *TreeExplainer* from the *shap* python library. The SA-Heart dataset was also used [13]. It was originally collected as part of an epidemiological investigation into coronary heart disease in South Africa. The dataset comprises clinical records from 462 individuals, with features including age, systolic blood pressure, cholesterol levels, tobacco use, family history of heart disease, and obesity-related metrics. A binary outcome variable indicating the presence or absence of coronary heart disease is also included (the prediction target). A RandomForest classifier was trained obtaining 0.76 and 0.8 F1-score for the test and train dataset, respectively.

Interaction tensor

In the context of tree-based models (*TreeExplainer*, *shap* python library), the first layer of interactions can be extracted, corresponding to a 3-dimensional tensor of shape (*n_samples*, *n_features*, *n_features*) where *n_samples* and *n_features* respectively represent the number of instances and features in the dataset. The diagonal entries of the tensor (across the last two dimensions) provide the main effect of each feature on the model prediction for each sample, without accounting for any interactions. The off-diagonal entries represent the interactions between pairs of features. Specifically, the entry at position (i, j, k) represents the interaction between feature j and feature k for sample i. These entries capture the combined effect of the two features that can not be attributed to their individual contributions. Such interactions are symmetric, meaning that for sample i, the interaction between feature j and feature k is identical to that between feature k and feature j.

Detection of the interaction trends

To inform about whether the interaction between a feature and another tends to amplify or attenuate the main effect of the feature itself, each interaction is multiplied by the sign of the SHAP contribution of the main effect. For instance, when the SHAP interaction value of *Glycohemoglobin* with *Triglycerides* (Fig 1B) is multiplied by the sign of the SHAP main effect of *Glycohemoglobin* (Fig 1A), it becomes clear that high values of *Triglycerides* tend to amplify the main (positive) effect of *Glycohemoglobin* (Spearman correlation coefficient $\rho = 0.78$, Fig 1C, 1D). In contrast, high values of *Cholesterol* and *ALT* tend to attenuate the main (positive) effects of *Glycohemoglobin* and *Cholesterol*, respectively ($\rho = -0.78$, $\rho = -0.82$, Fig 1A–1D).

Formally, if I is the 3-dimensional interaction tensor returned by *TreeExplainer*, a new 3-dimensional interaction tensor S can be computed so that:

$$S[i,k,j] = sign(I[i,j,j]) \cdot I[i,j,k] \quad (1)$$

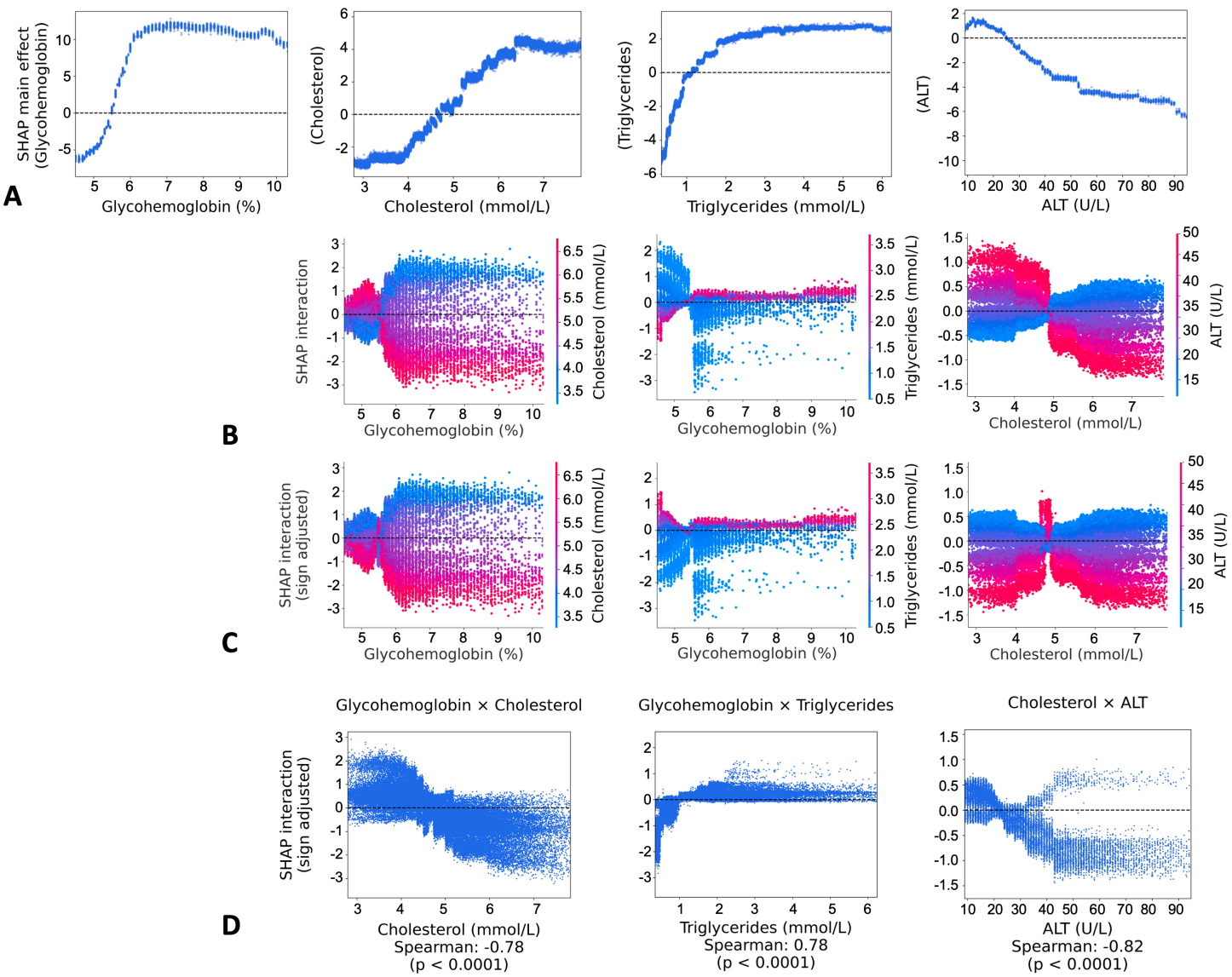

Fig 1. Representation of the treatment process to build the interaction graph. (A) Partial dependance plots displaying the SHAP main effect of each variable as a function of its raw value. Here, *Glycohemoglobin*, *Cholesterol* and *Triglycerides* have positive main effects while *ALT* has a negative main effect. **(B)** Partial dependence plots displaying SHAP interaction values for a pair of features as a function of the first feature's raw values (x-axis) and the second feature's raw values (color gradient). **(C)** Same as (C) but displaying SHAP interaction values after their multiplication with the sign of the first feature's SHAP main effect values from (A). For example, high values of *Cholesterol* tend to attenuate the main effect of *Glycohemoglobin* whereas high values of *Triglycerides* tend to amplify the main effect of *Glycohemoglobin*. **(D)** Detection of the interaction trends through Spearman's correlation coefficient. *Glycohemoglobin* × *Cholesterol* as well as *Cholesterol* × *ALT* will result in blue arrows on the graph whereas *Glycohemoglobin* × *Triglycerides* will result in a red arrow.

https://doi.org/10.1371/journal.pcsy.0000060.g001

where $I[i,j,j]$ is the main effect of feature j (diagonal entry). The resulting tensor has been transposed along its last 2 dimensions ($S[i,k,j]$ instead of $S[i,j,k]$) so that when feature j interacts with feature k, it is possible to see whether the values of feature k amplify or attenuate the main effect of feature j.

Consequently, computing the Spearman correlation coefficient allows to extract the direction and the strength of this association. The Spearman coefficient was chosen for its ability to detect monotonic non-linear relationships and its sign indicates the direction of

monotonicity. However, Pearson's correlation coefficient was also computed for its ability to detect monotonic linear relationships. A correlation threshold of –0.3 to 0.3 was chosen based on standard interpretation guidelines for Spearman's rank correlation coefficient, where values within this range are typically considered weak [14]; this allows us to visually highlight only moderate to strong monotonic relationships, while minimizing noise in the graphical representation.

Results

Each node of the graph (Fig 2) represents a biological variable, i.e. feature. Its size relates to the mean SHAP value of this feature's main effect while its color relates to the direction of the effect (i.e. red and blue for positive and negative correlation with predicted age, respectively). As an example for the prediction of physiological age, *Triglycerides*, *Cholesterol* and *Glycohemoglobin* are red nodes (positive correlation with predicted age) while *ALT* is a blue node (negative correlation with predicted age).

Each arrow from feature *k* to feature *j* indicates whether an increase in feature *k* amplifies or attenuates the main effect of feature *j*. High values of *Glycohemoglobin* amplify the previously identified positive main effect of *Triglycerides*, resulting in a red arrow on the graph but attenuate the previously identified positive main effect of *Cholesterol*, resulting in a blue arrow on the graph. High values of *Cholesterol* amplify the previously identified negative main effect of *ALT*, resulting in a red arrow on the graph. Additionally, a threshold can be set on the mean absolute SHAP values of nodes and arrows to allow only those of a certain importance to be displayed. In contrast, it is possible to observe situations in which the mean absolute SHAP value is high but the choice of color for the node or arrow is not straightforward. When Spearman's coefficient is between -0.3 and 0.3 (default value, parameter of the function [14]), the correlation is considered weak, the arrow is rendered dashed and particular attention must be paid to the visual analysis. In addition if Spearman's and Pearson's coefficients are of opposite signs, the corresponding node or arrow is rendered black as a warning. Such situations suggest a careful graphical analysis of the relationship is necessary.

The summary plot based on the SA-Heart dataset shows that the most important variables are *age*, *tobacco*, *famhist*, *ldl*, and *adiposity*. These also correspond to the most clearly interpretable explanations (Fig 3A). However, the explanations attributed to obesity may reflect interactions with other variables, particularly those related to metabolism (such as *ldl* or *adiposity*). Dependence plots do not clearly highlight any direct interaction between *adiposity* and *obesity*, whereas an interaction between *ldl* and *adiposity* appears more evident (Fig 3B, 3C). Nevertheless, capturing the full complexity of these interactions remains challenging. The interaction plot (Fig 3D) confirms the absence of a direct interaction between *adiposity* and *obesity*, but reveals strong positive synergistic interactions between *adiposity* and *ldl*, and between *famhist* and *obesity*.

Discussion and conclusion

This proposal shows the benefit of dissociating raw contributions from interaction contributions in additive explanation models in machine learning.

Compared to existing visualization methods for SHAP explanations, our proposed graph-based approach offers a novel way to capture the structure of feature interactions in a single, comprehensive figure. Unlike dependence plots that focus on pairwise relationships without contextualizing global effects, or heatmaps that are limited by scalability and visual clarity, our approach emphasizes a graphical and systemic representation of interaction dynamics. It provides an intuitive and information-rich view, particularly well-suited for biomedical

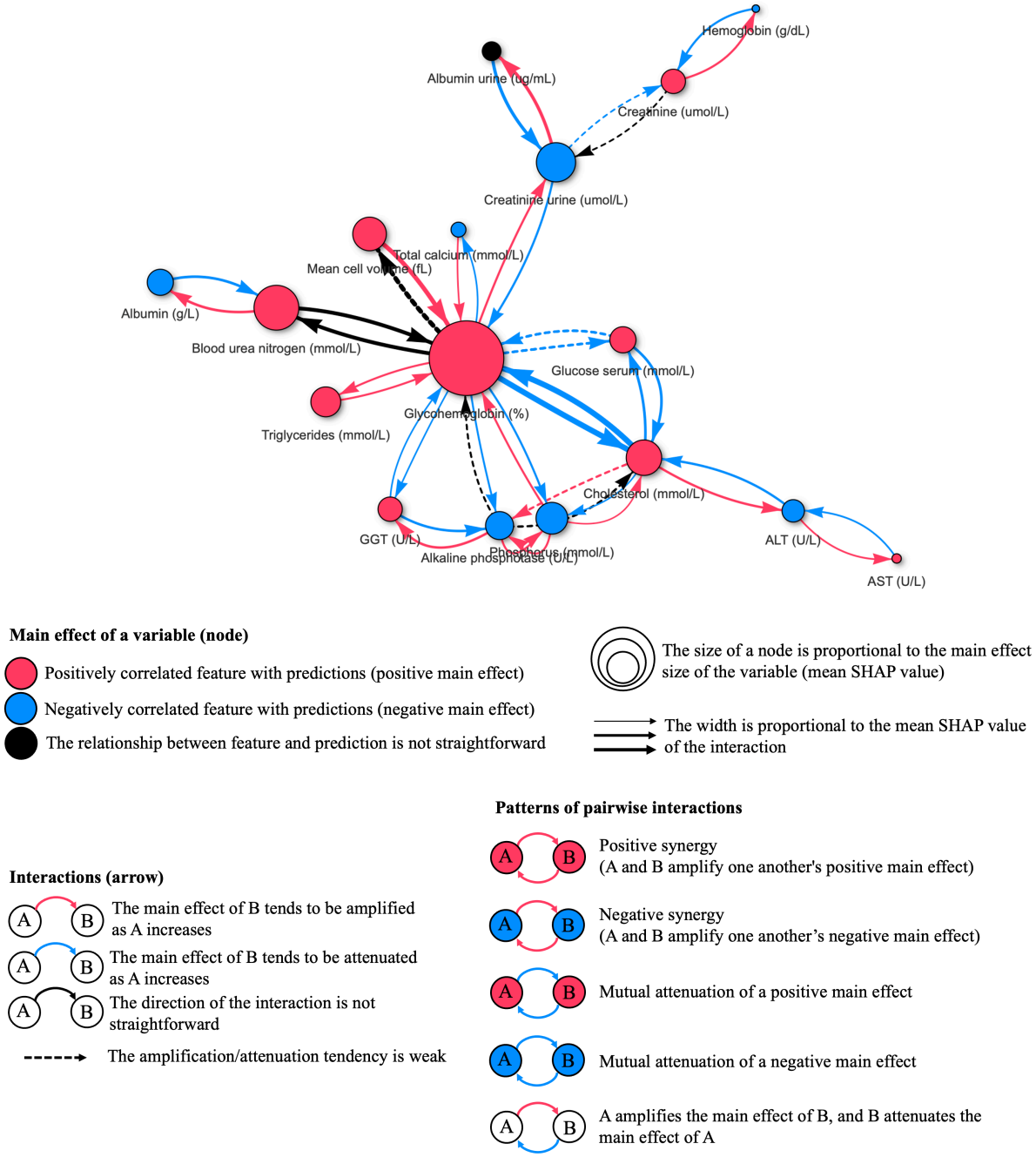

Fig 2. Representation of the interaction graph for Physiological Age. Each feature is a node whose size is proportional to the main effect and whose color indicates the sign of the correlation. Arrows between nodes are interaction values whose color indicates amplification or attenuation of the main effect of a feature when the other feature increases. When the correlation is weak (Spearman's correlation between −0.3 and 0.3), the arrow is rendered dashed. If Spearman's and Pearson's coefficients are of opposite signs, the corresponding node or arrow is rendered black as a warning. This graph displays the top 22% of feature interactions based on their mean absolute SHAP interaction values. This threshold was empirically chosen to balance readability and informational richness, and can be adjusted directly via a slider in the interactive graph to explore more or fewer interactions.

https://doi.org/10.1371/journal.pcsy.0000060.g002

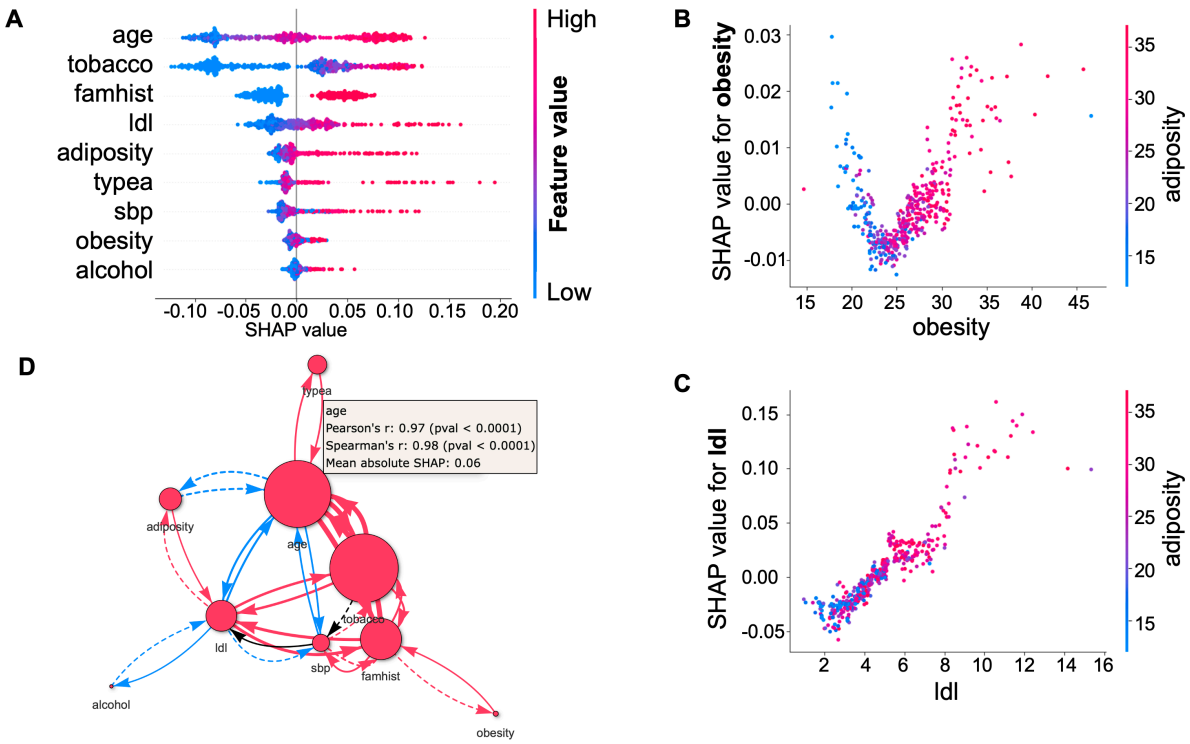

Fig 3. Analysis of feature contributions and interactions in the SA-Heart dataset. (A) SHAP summary plot of the top-10 most influential features in predicting coronary heart disease (CHD). All features, except *obesity*, exhibit a clearly positive contribution to CHD risk. (B) SHAP dependence plot illustrating the effect of the *obesity* feature on the model's predictions, with color indicating the corresponding *adipositiy* level. While *obesity* has a nonlinear impact on CHD risk (Spearman's r = 0.46, $p < 0.0001$), a clear interaction with *adiposity* is not observed: both low and high *adiposity* levels are associated with strong positive SHAP values, suggesting the absence of a consistent modifying effect. (C) SHAP dependence plot illustrating the effect of the *ldl* feature on the model's predictions (Spearman's r = 0.97, $p < 0.0001$), with color indicating the corresponding *adipositiy* level. The SHAP values increase linearly with *ldl* values, indicating a strong main effect. Additionally, higher *adiposity* values appear to amplify the effect of *ldl* on predictions, indicating a potential synergistic interaction between these two features. (D) Representation of the interaction graph with a interaction threshold of 0.2. The graph confirms the absence of a direct interaction between *obesity* and *adiposity*, but reveals significant positive interactions between *adiposity* and *ldl* (Spearman's r = 0.34, $p < 0.0001$), and between obesity and famhist (Spearman's r = 0.69, $p < 0.0001$).

https://doi.org/10.1371/journal.pcsy.0000060.g003

models where variables often interact in complex and subtle ways. While tools like SHAPash [8] and InterpretML [10] provide useful visual interfaces, they primarily focus on individual-level explanations or variable importance, and do not directly address interaction dynamics. Our approach complements these tools by enabling a higher-level systemic understanding of model behavior through interaction patterns. While SHAP-IQ [11] is useful for identifying and quantifying top interactions through scoring metrics, it lacks an integrative visual layout that reveals how multiple features collectively influence one another. Our method leverages the complete SHAP interaction tensor and computes signed correlations to uncover whether a feature tends to amplify or attenuate the effect of another. This results in a directed graph where both interaction strength and directionality are encoded, enabling the discovery of higher-order patterns such as mutual attenuation or dominant influences. Such representations are particularly suited to biomedical systems, where understanding the interplay between variables—rather than isolated pairwise effects—is often key to interpreting complex physiological mechanisms.

Different approaches to obtain interaction values are possible such as Accumulated Local Effects (ALE) plots [15] or Friedman's H-statistic [16]. Accumulated Local Effects (ALE)

plots overcome key limitations of partial dependence plots (PDP) by isolating feature effects through localized prediction differences within feature intervals, avoiding PDP's unrealistic extrapolations from marginal distributions when features are correlated. For interaction analysis, Friedman's H-statistic quantifies interaction strength by comparing joint partial dependence effects to individual feature contributions, using decomposition principles to measure both pairwise and total interaction strengths. However, these methods typically offer limited resolution at the instance level and do not distinguish between amplifying or attenuating effects in a directional manner. In contrast, our method explicitly leverages SHAP interaction values to construct a directed graph that captures both the strength and polarity of interactions, enabling a more interpretable and biologically meaningful synthesis. A combination of ALE and our proposed method could yield especially informative and complementary insights.

Even if this visualization only highlights monotonic trends, it may be possible to supplement statistical evaluations with correlation statistics adapted to non-monotonic relationships [17], although this type of relationship makes interpretation extremely difficult. In addition, although the present work is based on the explanations provided by *SHAP* (*TreeExplainer*), it applies to all local explanation techniques that make it possible to dissociate the main effect from that of interactions such as complete or coalitionnal methods [6]. This optimized visualization reveals the central role of *Glycohemoglobin*, that not only directly strongly impacts the prediction of physiological age but is in interaction with a large number of features. The patterns of pairwise interactions demonstrate a dominant effect of *Glycohemoglobin* for several interactions as well a positive synergy with *Triglycerides* and a mutual attenuation with *Cholesterol*. Such mutual attenuation may reveal homeostatic regulatory mechanisms. The same applies to the prediction of coronary heart diseases, where the graphical representation clearly illustrates the complexity of interactions between the various risk factors. In particular, it highlights that *adiposity* and *obesity* capture different physiological phenomena and interact with distinct sets of variables.

While this representation effectively captures the dynamics of interactions, it increases the dimensionality of the information to be analyzed (raw data, individual variable contributions, interaction contributions, and interaction dynamics). As a result, it becomes necessary to reduce the exploration space and to give the user control over this exploration. We propose here a simple filtering approach based on the top percentage of mean absolute SHAP interaction values, retaining only the most relevant feature pairs; this percentage can be adjusted interactively via a slider on the graph. For future applications involving larger feature sets, the framework can be extended to incorporate dimensionality reduction techniques such as UMAP or PCA, or unsupervised clustering methods like hierarchical clustering, to identify coherent groups of interacting features while preserving global interpretability.

By incorporating interaction effects, the explanation space evolves into a higher-dimensional, information-rich representation, enabling deeper insight into the model's decision-making process and supporting knowledge discovery. More than just a way of rendering information, such a methodology could be applied to many biomedical contexts, including RNA-sequencing, enabling us to improve our understanding of pathophysiology.

Acknowledgments

Availability of source code and requirements

- Project name: SHAP-Interactions
- Project home page:
  https://pypi.org/project/shapinteractions/

- Operating system(s): Platform independent
- Programming language: Python
- License: GNU GPL

Author contributions

Conceptualization: Julien Aligon, Louis Casteilla, Paul Monsarrat.

Data curation: Miguel Thomas, Emmanuel Doumard.

Formal analysis: Emmanuel Doumard, Paul Monsarrat.

Funding acquisition: Julien Aligon, Paul Monsarrat.

Investigation: Felix Furger, Miguel Thomas, Paul Monsarrat.

Methodology: Felix Furger, Julien Aligon, Cyrille Delpierre, Louis Casteilla, Paul Monsarrat.

Project administration: Paul Monsarrat.

Resources: Julien Aligon.

Software: Felix Furger, Miguel Thomas, Emmanuel Doumard, Paul Monsarrat.

Supervision: Louis Casteilla, Paul Monsarrat.

Validation: Cyrille Delpierre, Paul Monsarrat.

Visualization: Felix Furger, Emmanuel Doumard, Paul Monsarrat.

Writing – original draft: Felix Furger, Miguel Thomas, Emmanuel Doumard.

Writing – review & editing: Julien Aligon, Cyrille Delpierre, Louis Casteilla, Paul Monsarrat.

References

1. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019;25(1):44–56. https://doi.org/10.1038/s41591-018-0300-7 PMID: 30617339
2. Beam AL, Kohane IS. Big data and machine learning in health care. JAMA. 2018;319(13):1317–8. https://doi.org/10.1001/jama.2017.18391 PMID: 29532063
3. Muehlematter UJ, Daniore P, Vokinger KN. Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (2015-20): a comparative analysis. Lancet Digit Health. 2021;3(3):e195–203. https://doi.org/10.1016/S2589-7500(20)30292-2 PMID: 33478929
4. He J, Baxter SL, Xu J, Xu J, Zhou X, Zhang K. The practical implementation of artificial intelligence technologies in medicine. Nat Med. 2019;25(1):30–6. https://doi.org/10.1038/s41591-018-0307-0 PMID: 30617336
5. Yoon CH, Torrance R, Scheinerman N. Machine learning in medicine: should the pursuit of enhanced interpretability be abandoned?. J Med Ethics. 2022;48(9):581–5. https://doi.org/10.1136/medethics-2020-107102 PMID: 34006600
6. Doumard E, Aligon J, Escriva E, Excoffier J-B, Monsarrat P, Soulé-Dupuy C. A quantitative approach for the comparison of additive local explanation methods. Information Systems. 2023;114:102162. https://doi.org/10.1016/j.is.2022.102162
7. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017. p. 4768–77.
8. Shapash. User-friendly explainability and interpretability to develop reliable and transparent machine learning models. 2025. https://maif.github.io/shapash/
9. Baniecki H, Kretowicz W, Piatyszek P, Wisniewski J, Biecek P. dalex: responsible machine learning with interactive explainability and fairness in python. Journal of Machine Learning Research. 2021;22(214):1–7.
10. InterpretML. Fit interpretable models. Explain blackbox machine learning. 2025. https://github.com/interpretml/interpret/

11. Muschalik M, Baniecki H, Fumagalli F, Kolpaczki P, Hammer B, Hüllermeier E. shapiq: shapley interactions for machine learning. In: The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track; 2024.
12. Bernard D, Doumard E, Ader I, Kemoun P, Pagès J-C, Galinier A, et al. Explainable machine learning framework to predict personalized physiological aging. Aging Cell. 2023;22(8):e13872. https://doi.org/10.1111/acel.13872 PMID: 37300327
13. Rossouw JE, Du Plessis JP, Benadé AJ, Jordaan PC, Kotzé JP, Jooste PL, et al. Coronary risk factor screening in three rural communities. The CORIS baseline study. S Afr Med J. 1983;64(12):430–6. PMID: 6623218
14. Hinkle DE, Wiersma W, Jurs S. Applied statistics for the behavioural sciences. 5th ed. Houghton Mifflin (Academic); 2002.
15. Apley DW, Zhu J. Visualizing the effects of predictor variables in black box supervised learning models. Journal of the Royal Statistical Society Series B: Statistical Methodology. 2020;82(4):1059–86. https://doi.org/10.1111/rssb.12377
16. Friedman JH, Popescu BE. Predictive learning via rule ensembles. Ann Appl Stat. 2008;2(3). https://doi.org/10.1214/07-aoas148
17. Wang H, Aligon J, May J, Doumard E, Labroche N, Delpierre C, et al. Discernibility in explanations: Designing more acceptable and meaningful machine learning models for medicine. Comput Struct Biotechnol J. 2025;27:1800–8. https://doi.org/10.1016/j.csbj.2025.04.021 PMID: 40458636