DeepDiff-SHAP: Interpretable deep learning for subgroup-specific causal hypothesis generation using conditional SHAP

Aditya Sriram[1], Soyeon Kim[2], Joseph A Carcillo[2], Hyun Jung Park[1,*]

[1]*Department of Human Genetics, University of Pittsburgh, Pittsburgh, PA, USA*
[2]*Department of Pediatrics, University of Pittsburgh, Pittsburgh, PA, USA*
[*]*E-mail: hyp15@pitt.edu*

Precision medicine aims to tailor healthcare strategies to individual differences in genetic, clinical, and environmental factors. However, identifying subgroup-specific causal relationships in complex biomedical data remains a major challenge, especially when standard causal inference methods average over population heterogeneity. We introduce DeepDiff-SHAP, a novel framework that combines regression-based and deep learning-based differential causal inference to detect changes in causal relationships across patient subgroups. DeepDiff-SHAP integrates conditional SHapley Additive exPlanations (SHAP) to estimate conditional dependencies and perform nonlinear differential causal inference in a principled, interpretable manner. Applying DeepDiff-SHAP to two population-scale datasets, the CDC Diabetes Health Indicators Dataset and a UK Biobank sepsis cohort stratified by hypertension status, we identified clinically meaningful and subgroup-specific causal changes in relationships between features across the datasets including age, general health, alkaline phosphatase, and cholesterol. Our results reinforce the idea that deep learning enhances sensitivity to complex interaction patterns overlooked by linear models, providing new biological insights into disease progression and comorbidity-specific risk mechanisms. DeepDiff-SHAP offers a scalable and interpretable solution to uncover individualized causal pathways, advancing the goal of truly personalized medicine.

Keywords: Differential Causal Inference; Deep Learning; Precision Medicine; Shapley Additive exPlanations (SHAP).

1. Introduction

Precision medicine is transforming biomedical research and clinical care by shifting the focus from one-size-fits-all treatments to strategies tailored to individual differences in genetic, environmental, and lifestyle factors. As precision medicine continues to gain traction, there is a growing need for analytical methods that can identify subgroup-specific risk factors that either influence disease risk or therapeutic responses differently across distinct populations, or differently across multiple states within the same population. Examples include genetic variants such as *rs11673407* in the fucosyltransferase 3 gene (*FUT3*) elevating cardiovascular risk in men but not in women[1], or nine potentially protective and 25 harmful metabolic biomarkers predicting future incidence of type 2 diabetes[2]. Traditional causal inference frameworks, however, typically estimate average treatment or exposure effects across a global population grouping[3-5]. As an example, CausalMGM is a well-established mixed graphical model method for inferring causal relationships from observational data that may include multiple data types[6]. However, since it runs on a single aggregated cohort, it can obscure possible nuanced, group-specific mechanisms and lead to ineffective or even harmful interventions in underrepresented subgroups within the larger data group.

Recently, a method proposed by Belyaeva et al. called Differential Causal Inference (DCI), established a principled approach to detect differences in causal effects across groups by directly comparing the strength of variable-outcome relationships in one subgroup versus another[7]. DCI is methodologically distinct from simply performing causal inference separately in two groups and comparing the results. In a standard two-group approach, causal models are estimated independently for each group (e.g., diseased and healthy), and differences in effect estimates are compared post hoc. However, this approach does not account for estimation variance or the statistical significance of the differences, potentially leading to spurious findings or uncertainty in drawing conclusions. In contrast, DCI directly models and tests the difference in causal effects between groups as the primary quantity of interest within a unified framework that allows for more robust inference. This enables the identification of subgroup-specific causal mechanisms while controlling for variability and potential biases. Furthermore, the naive approach of fitting models separately in subgroups, especially in high-dimensional settings, can lead to unstable estimates due to small sample sizes or overfitting. DCI's framework borrows strength across groups through joint modeling or shared representations, improving estimation accuracy and interpretability. This enables researchers and clinicians to uncover risk factors that are uniquely relevant to particular population subgroups, such as non-responders to immunotherapy[8] or patients with treatment-resistant depression[9], and in turn advance the development of more precise and targeted interventions.

Despite its advantage as being the only current differential causal inference method, DCI is based on regression-based framework and thus limited in its ability to capture the sophisticated mechanisms underlying disease heterogeneity. Disease heterogeneity arises from complex, multilayered biological processes that involve nonlinear interactions among genetic, epigenetic, and environmental factors. Regression models, which rely on additive and linear assumptions, are insufficient in capturing these involved dependencies, potentially overlooking key disease-driving mechanisms.

To model complex biological and clinical systems, which are governed by intricate, nonlinear interactions among molecular signals (e.g., gene expression, DNA methylation)[10,11], environmental

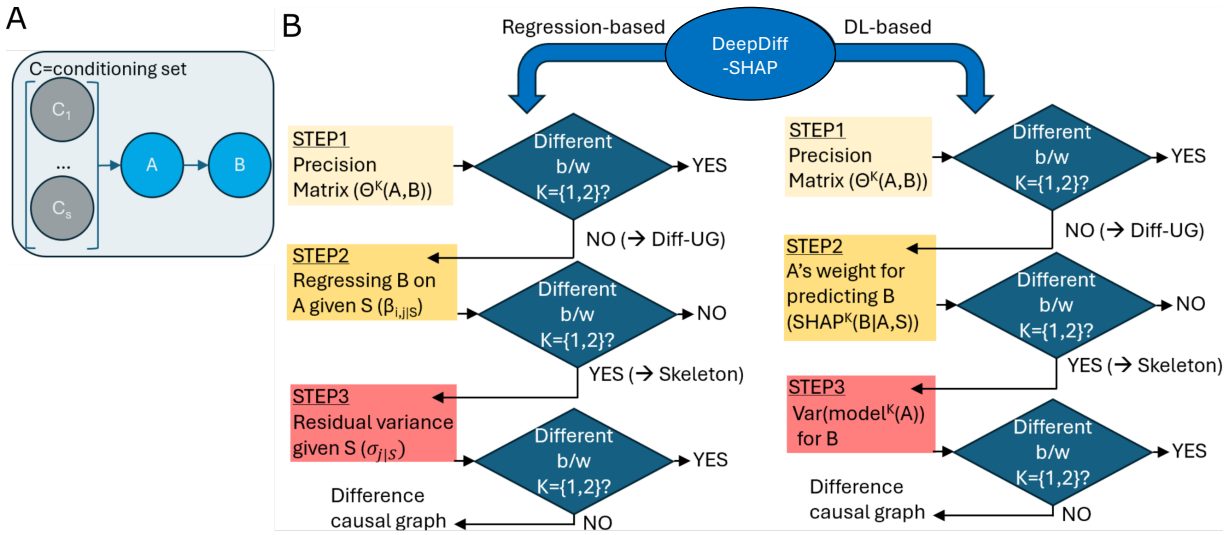

Figure 1. (A) Causal relationship from A to B to test for the difference based on the conditioning set expressed as $C = \{C1, \ldots, Cs\}$. (B) Algorithm of DeepDiff-SHAP with the regression and the deep learning components.

exposures (e.g., smoking, pollution)[12,13], and clinical variables (e.g., comorbidities, treatment history)[14,15], we will extend regression-based DCI using deep learning (DL). With its multilayered architecture and nonlinear activation functions, DL can effectively learn hierarchical feature representations that capture subtle, high-order dependencies between inputs. It has already demonstrated success in a variety of biomedical tasks including disease risk prediction[16,17] and image-based diagnostics[18,19]. This makes deep learning particularly well-suited for extending DCI for contexts in which causal effects are likely to vary not just in magnitude, but in functional form, across subgroups.

Specifically, for a framework outlining a causal association from input feature A to outcome B with conditioning on set C (**Fig. 1A**), we will extend the established regression-based three step approach of DCI[7] (**Fig. 1B**) by using DL architecture. While this approach evaluates model parameters for the conditional probability of each variable pair between conditions, performing these steps requires conditional probabilities, e.g., $P(A|B,C)$, where C is a set of conditioning nodes. In the regression setting, such conditional distributions have closed-form expressions based on covariances and are thus tractable. However, in DL settings, these conditional distributions become complex and high-dimensional. To address this, we propose a novel deep learning framework, **DeepDiff-SHAP**, that incorporates advanced SHapley Additive exPlanations (SHAP)[20] to quantify differential conditional dependence between variables (**Fig. 1C**). SHAP has been adapted to estimate conditional expectations in supervised learning[21], but its use for computing conditional dependencies between input variables, as required for formal causal inference, remains limited and underexplored. By adapting SHAP to contrast feature contributions of certain variables, while conditioning on the shared, high-dimensional covariate space that exists in most biomedical data, our method enables robust and interpretable identification of variables whose causal effects differ across groups. This approach directly targets the core objective of DCI and offers a scalable, principled solution for uncovering subgroup-specific causal mechanisms in complex biomedical data. Below, we will first formalize the three-step workflow of DeepDiff-SHAP: (1) screen for causal edges that change between two states by contrasting precision matrices; (2) prune edges that may appear variant but do not actually differ, using a SHAP-based test of whether one variable's influence on another still differs after conditioning on nearby variables; and (3) orient the remaining edges by checking which putative parent sets keep prediction residuals stable across states using lightweight neural networks. We apply DeepDiff-SHAP to two case studies: (1) CDC Diabetes survey data, and (2) a UK Biobank cohort of sepsis patients stratified by hypertension status, both representing disease groups with distinct etiologies and risk structures for which we discuss key empirical findings. Finally, we provide practical guidance on aligning comparison groups with the intended causal effect modifiers and conclude with a transparent discussion of limitations and planned methodological improvements.

2. Introduction

We introduce DeepDiff-SHAP as a principled, statistically sound three-algorithm approach rooted in regression and DL frameworks for causal structure changes between two states (**Fig. 1B**). DeepDiff-SHAP works without having to fully reconstruct each underlying network in a dataset. Step 1 involves identifying candidate nodes and edges where the dependency structure differs between states, based on changes in the precision matrix; in this case, the precision matrix is the pseudoinverse of the empirical covariance matrix. Since structural changes in a causal graph often result in shifts in the precision matrix, this step ensures a stable, high recall starting point for further

evaluation and refinement. In Step 2, we test whether the strength of variable dependency relationships (i.e., how much one variable predicts another after conditioning on a third variable in the undirected graph) remains stable across the two states. If a relationship does not change, it is pruned from the candidate set, thus removing possible false positives identified by Step 1. Finally, in Step 3, we compute differences in residual (unexplained) variances calculated via deep neural networks (DNNs) to infer directionality between nodes; if there is statistical evidence of DNN residual invariance when conditioning on a set of nodes, it suggests that the node set contains the true causal parents related to the outcome.

2.1. *Estimation of the difference undirected graph*

To identify pairwise changes in conditional dependence structure between two states, we begin by estimating a **difference undirected graph (Δ-UG)** that aims to identify statistically significant evidence of interactions that vary between the two state-separated data groups. Let $X^{(1)} \in \mathbb{R}^{n_1 \times p}$ and $X^{(2)} \in \mathbb{R}^{n_2 \times p}$ denote independent data groups from states 1 and 2, respectively. Each dataset is denoted by state-specific precision matrices $\Theta^{(1)} = \Sigma^{(1)^{-1}}$ and $\Theta^{(2)} = \Sigma^{(2)^{-1}}$, where $\Sigma^{(k)}$ is the covariance matrix of state k.

We implement a constraint-based approach that computes an edge-specific test statistic for each pair of variables (i, j) based on their estimated precision matrix entries. This step is adapted from the framework first formalized by Belyaeva et al[22]. Specifically, we estimate $\widehat{\Theta}^{(1)}$ and $\widehat{\Theta}^{(2)}$ using the Moore–Penrose pseudoinverse of the empirical covariance matrices computed from $X^{(1)}$ and $X^{(2)}$, respectively. The test statistic for each pair (i, j) is defined as:

$$\widehat{Q}_{ij} \coloneqq \frac{\left(\widehat{\Theta}_{ij}^{(1)} - \widehat{\Theta}_{ij}^{(2)}\right)^2}{\left(\frac{\widehat{\Theta}_{ii}^{(1)}\widehat{\Theta}_{jj}^{(1)} + (\widehat{\Theta}_{ij}^{(1)})^2}{n_1} + \frac{\widehat{\Theta}_{ii}^{(2)}\widehat{\Theta}_{jj}^{(2)} + (\widehat{\Theta}_{ij}^{(2)})^2}{n_2}\right)}.$$

This statistic quantifies the squared difference in partial correlations between the two states, scaled by their estimated variances. Under the null hypothesis $H_0: \Theta_{ij}^{(1)} = \Theta_{ij}^{(2)}$, the statistic $\widehat{Q}_{ij}$ asymptotically follows a noncentral F-distribution[23,24]:

$$\widehat{Q}_{ij} \sim F(1, \nu), \quad \text{where} \quad \nu = n_1 + n_2 - 2p + 2.$$

We compute p-values from this distribution using the cumulative noncentral F distribution function and define a significance threshold $\alpha \in [0,1]$. The difference undirected graph Δ is then constructed as:

$$\mathcal{E}_\Delta \coloneqq \left\{\{i, j\} \;\middle|\; p_{ij} \leq \alpha\right\},$$

where p_{ij} is the p-value corresponding to $\widehat{Q}_{ij}$. To reduce the downstream hypothesis testing burden in skeleton discovery and edge orientation, we define the set of conditioning nodes as:

$$\mathcal{C} \coloneqq \{i \mid \exists j \text{ such that } \{i, j\} \in \mathcal{E}_\Delta\}.$$

We limit the nodes included in the conditioning set to the nodes involved in the edges of the difference undirected graph, Δ-UG. We block the inclusion of any additional nodes (i.e. marginal or conditional distributions differing across states) leading to node inclusion in the conditioning set, ensuring a very strict and regulated edge inclusion step.

2.2. *Skeleton discovery via SHAP-based conditional invariance testing*

In the second step of our method, we further prune the initial undirected difference graph by testing for conditional invariance of cross-state feature dependencies. Specifically, for each edge (i, j) in the initial undirected difference graph, we assess whether the importance of feature i for predicting feature j remains invariant across states, and vice versa, after conditioning on subsets of features part of the conditioning set S.

We model each feature i as a function of its potential parent j and a conditioning set $S \subseteq \mathcal{C}\{i, j\}$, where $\mathcal{C}$ denotes the set of conditioning nodes obtained from Step 1. For each direction, we train two multilayer perceptron regressors $f_i^{(1)}$ and $f_i^{(2)}$ on the two datasets $X^{(1)}$ and $X^{(2)}$ to predict X_i using X_j and S.

To assess the contribution of X_j to the prediction of X_i in each state, we compute conditional SHAP values using KernelSHAP with a fixed background distribution that isolates the effect of X_j given S[25]. Specifically, we calculate:

$$\phi_{j|S}^{(k)} = \mathrm{SHAP}\left(X_j \,\middle|\, S; f_i^{(k)}, X^{(k)}\right),$$

where $k \in \{1,2\}$ and $\phi_{j|S}^{(k)}$ denotes the absolute SHAP values across test samples. These values are then compared between states using a normalized squared difference statistic:

$$T_{ij|S} := \frac{(\mu_1 - \mu_2)^2}{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}},$$

where μ_k and σ_k^2 are the mean and variance of the SHAP values in state k, and n_k is the number of SHAP samples. A two-sided p-value is derived from the noncentral F-distribution with degrees of freedom $\nu = n_1 + n_2 - 2 - 2|S|$.

For each ordered pair ($i \leftarrow j$ and $j \leftarrow i$) we test for SHAP heterogeneity across states conditional on the set of conditioning nodes. When $p > \alpha$, we fail to reject the null hypothesis of conditional invariance across states and remove the corresponding edge from the undirected difference graph. When $p < \alpha$, we reject the null hypothesis of conditional invariance, and we include the edge. This process is repeated across all conditioning sets up to a specified maximum size $r_{\max}$. The remaining edges after this pruning step make up the edge difference skeleton used in the subsequent direction-orientation step for the leftover edges.

This SHAP-based invariance testing allows for nonlinear, model-flexible detection of asymmetric changes in feature relationships, extending the regression-based conditional independence tests to deep neural network models with a more structurally sound framework.

2.3. *Edge orientation via invariance testing*

In step 3, we orient edges obtained in step 2 (the pruned edge difference skeleton) to end up with a directed graph that captures differences in functional dependencies across the two states. We assume that for any node j, if the conditional variance of X_j given a set of likely parents S is invariant across states, then S represents a valid set of parents for j. This principle is rooted in the theory that the functional form of the conditional distribution $P(X_j \mid X_S)$ should remain stable under invariance[26].

For each node j in the graph, we test candidate parent sets $S \subseteq \mathcal{C}\{j\}$ of size $k = 1$, where $\mathcal{C}$ is the set of conditioning nodes identified from the pruned edge difference skeleton from algorithm 2. We train DNNs to regress X_j on X_S separately in each state using a two-layer multilayer perceptron (MLP). The residual variance is estimated as:

$$\hat{\sigma}_{j|S}^{2\,(k)} \coloneqq \mathrm{Var}\left(X_j^{(k)} - f_j^{(k)}\left(X_S^{(k)}\right)\right),$$

where $f_j^{(k)}$ denotes the fitted DNN regressor in state $k \in \{1,2\}$, and $\hat{\sigma}^2$ is computed on the training data. To assess whether the conditional variance differs significantly between states, we compute a two-sided test statistic based on the ratio of residual variances:

$$T_{j|S} \coloneqq \frac{\hat{\sigma}_{j|S}^{2\,(1)}}{\hat{\sigma}_{j|S}^{2\,(2)}},$$

with a p-value computed using the noncentral F-distribution with degrees of freedom $(n_1 - |S|, n_2 - |S|)$:

$$p = 2 \cdot \min\left(F_{\mathrm{cdf}}(T_{j|S}), 1 - F_{\mathrm{cdf}}(T_{j|S})\right).$$

If the p-value exceeds a specified threshold α, we fail to reject the null hypothesis and conclude that the conditional variance is invariant, therefore accepting S as the parent set for j. The directed edges $i \rightarrow j$ for all $i \in S$ are added to the graph. We perform additional cycle and contradiction checks using transitive closure on the directed graph to prevent invalid orientations.

For any edges that remain unoriented after this test, we apply graph traversal rules to resolve directionality wherever a consistent path structure allows. For example, for any set of nodes in which $i \rightarrow node_1 \rightarrow node_x \rightarrow j$, we orient $i \rightarrow j$. This rationale is formalized by Meek and reviewed by Colombo et al[27,28]. The final output is a directed adjacency matrix, a log of all the orientation decisions, and the set of edges that could not be oriented by algorithm 3's invariance testing.

This DNN-based residual variance orientation strategy expands on the original DCI steps defined for regression-based models and leverages DL via a DNN framework to capture potentially nonlinear predictive structure, and tests whether this structure is preserved across states; this allows for stronger performance of causal inference in a model-independent way[22]. Our model eliminates the assumption of linear-Gaussian data by utilizing DNN-based predictions for variable dependencies.

For results mentioned in this paper, DeepDiff-SHAP was initialized with the following parameters: $\alpha_{ug} = 0.005, \alpha_{skeleton} = 0.3, \alpha_{orient} = 0.001$ corresponding to the threshold levels for each of the three steps of the algorithm. Conditioning set size was set to 1 (maximum set size is 2, range from 0 to 2; a higher conditioning set size leads to sparser causal network graphs).

2.4. Computational feasibility of conditional-based SHAP

While standard KernelSHAP is known to suffer from high computational cost, it is important to note that conditional SHAP can reduce computational cost significantly. In the cost calculation, computational expense in KernelSHAP scales as $2^{(p)}$ model evaluations, where p is the number of features. On the other hand, instead of evaluating all subsets of non-target features, conditional SHAP computes the contribution of a predictor j given a limited conditioning set S, integrating over the distribution of the remaining features via $p(X_{S\setminus\{j\}} \mid X_S)$. As a result, the model evaluation space collapses from size 2^p to $2^{|S|}$ (the subsets of $\{j\} \cup S$). Since our algorithm restricts $| S | \leq r_{\max}$, we can significantly reduce the number of model evaluations. Reusing a shared conditional background distribution across all test instances further reduces per-sample computational load.

3. Results

3.1. *DeepDiff-SHAP reveals nonlinear subgroup-specific causal structures in diabetes populations*

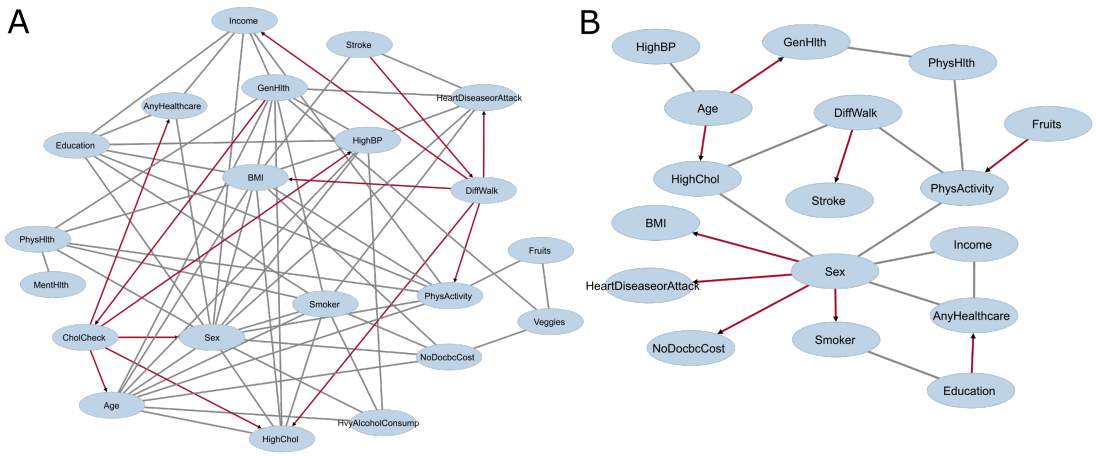

Figure 2. DCI graph on the diabetes data as identified by (A) the regression-based and (B) deep-learning-based DeepDiff-SHAP. Gray lines represent the association changes that are not attributed as differential causal relationship and red lines represent the differential causal relationships.

To investigate changes in causal relationships associated with chronic disease, we ran the regression-based DCI and DL-based DeepDiff-SHAP on the 2014 Centers for Disease Control and Prevention (CDC) Diabetes Health Indicators Dataset[29]. This dataset includes health status, behavioral, and access-to-care survey data for 253,680 individuals, among whom 39,977 were diagnosed as diabetic or prediabetic and 213,703 were not diagnosed with diabetes. Using the regression-based module of DeepDiff-SHAP, which follows the original DCI framework, we detected 68 feature pairs with differential associations between the diabetic and non-diabetic groups. Many of these association changes involved well-known demographic confounders such as age and sex, which influence a wide range of lifestyle and health indicators. For instance, age was differentially associated with features like general health (labeled "GenHlth"), stroke, smoking status (labeled "Smoker"), and access to health care (labeled "AnyHealthcare"), while sex was

associated with differences in income, smoking status, and BMI (**Fig. 2A**). 12 of the 68 differential associations (17.64%) were attributed to underlying causal relationship changes by the regression-based model. Notably, among participants categorized as diabetic or prediabetic, worse general health exerted a stronger causal effect on obtaining a cholesterol check within 5 years (labeled "CholCheck"), and, in turn, the check had a causal link with individuals who had high-cholesterol (labeled "HighChol") and high-blood pressure (HighBP); these relationships were not observed in the non-diabetic patient group. However, only 2 (16.66%) of the 12 findings involved age or sex as causal variables, despite their central role in shaping health outcomes. While it is clinically plausible that age impacts general health status differently in people with and without diabetes, this causal difference was not detected by the regression model. In contrast, the deep learning module of DeepDiff-SHAP (**Fig. 2B**) identified this expected causal difference[30] from age to general health, as well as a differential causal effect of age on high cholesterol. Similarly, while the regression-based model did not attribute any of the sex-related association changes, such as those from sex to heart disease or heart attack (labeled "HeartDiseaseAttack"), smoking status, and BMI, to changes in causal relationships, the DL-based approach successfully identified all these as differential causal effects between the diabetic/pre-diabetic and non-diabetic groups.

We note the importance of these particular causal relationships in the DeepDiff-SHAP network graph, age to high cholesterol and sex to coronary heart disease or myocardial infarction, both which have been previously mentioned in the context of type 2 diabetes being associated with women having a greater risk of cardiovascular disease, as well as studies implicating earlier cardiovascular disease events in women with T2D compared to men[31-33]. These findings suggest that deep learning-based causal inference can uncover subtle and nonlinear changes in causal structure that may be missed by linear models. Notably, the regression-based DCI model identified some relationships that appear potentially misleading. As an example, regression-based DCI attributed a directional causal difference in the relationship from stroke to difficulty walking, which is more plausibly explained in the reverse direction or mediated by baseline functional impairments in diabetic individuals. Altogether, these results both highlight the limitations of traditional regression-based causal inference in detecting meaningful shifts in causal mechanisms between disease subpopulations as well as showcase how our deep learning framework enables more nuanced detection of differential causal structures, supporting its utility in understanding the biological and behavioral heterogeneity of chronic diseases such as diabetes.

3.2. *Uncovering comorbidity-specific mechanisms in sepsis through causal inference*

To investigate how chronic comorbidities modulate causal relationships in sepsis, we applied DeepDiff-SHAP (**Fig. 1B**) to a subset of the UK Biobank patient database comprising 3,181 individuals diagnosed with sepsis, stratified by hypertension status (hypertension: n = 2,669; no hypertension: n = 512). Sepsis remains a leading cause of morbidity and mortality in adults and children, yet its clinical progression is strongly influenced by pre-existing conditions such as hypertension, a factor often overlooked in risk modeling[34]. We identified sepsis cases using ICD-10 codes (e.g., A40, A41, B37.7, O85), capturing a range of septicemia and related conditions. From the UK Biobank, we selected 42 variables spanning domains of cardiometabolic health, renal and liver function, inflammation, hormones, blood pressure, and pulmonary status. Using DeepDiff-SHAP's regression-based module, we identified 18 associations that differed between hypertensive and non-hypertensive sepsis patients, of which 8 (44.44%) were attributed to shifts in causal relationships (**Fig. 3A**). Notably, these included altered causal links including from urate to SHBG

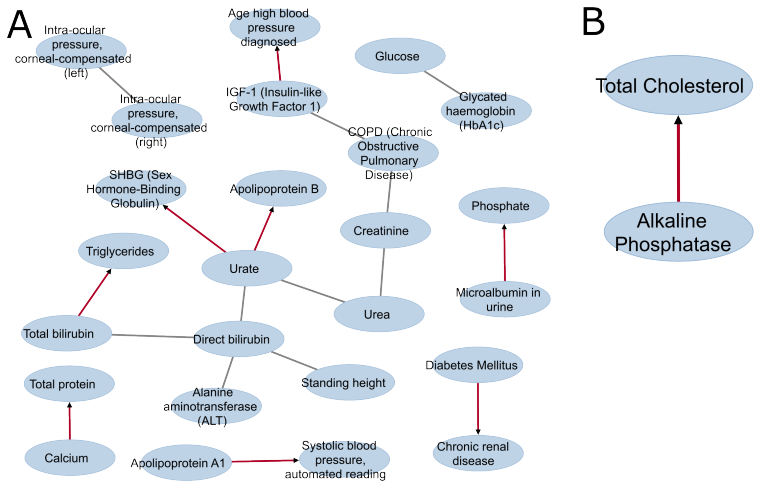

Figure 3. Differential causal graphs on UK Biobank sepsis data as identified by (A) the regression-based and (B) deep-learning-based DeepDiff-SHAP. Gray lines represent the association changes that are not attributed as differential causal relationship and red lines represent the differential causal relationships.

and apolipoprotein B, total bilirubin to triglycerides, calcium to total protein, apolipoprotein A1 to systolic blood pressure, and IGF-1 to age at hypertension diagnosis, suggesting distinct physiological pathways influenced by vascular stress. Importantly, the deep learning component of DeepDiff-SHAP uncovered an additional causal shift from alkaline phosphatase to total cholesterol that was not detectable via the regression-based model (**Fig. 3B**), highlighting the added value of incorporating nonlinear modeling.

Together, these findings demonstrate that DeepDiff-SHAP enables a comprehensive discovery of subgroup-specific causal mechanisms in sepsis by building on regression-driven DCI with a deep learning framework.

3.3. *Benchmarking DeepDiff-SHAP*

3.3.1. *Benchmarking DeepDiff-SHAP against regression-based DCI: enhanced sensitivity to nonlinear causal differences*

To benchmark performance, we compared DeepDiff-SHAP against DCI, the only existing differential causal inference method currently available. Importantly, the regression-based component of DeepDiff-SHAP is mathematically equivalent to this original DCI method, as both frameworks follow the same three-step procedure: identifying structural differences, testing conditional invariance, and estimating directionality based on residual variance. Therefore, any observed differences in results between DeepDiff-SHAP's deep learning (DL) module and the regression-based DCI can be attributed to the modeling framework itself, rather than the procedural design. To assess relative sensitivity to causal structure changes, we compared the number of association differences that were ultimately attributed to changes in causal strength (i.e., differential causal degrees). Since the comparison holds the inference pipeline constant while varying only the functional form (linear vs. nonlinear approach), this metric offers a controlled and clear evaluation of how effectively each approach detects meaningful causal differences. Between the diabetes (pre-

diabetic and diabetic diagnosis) and healthy population groups in the CDC Diabetes dataset, the original DCI method identified 68 association changes, but only 12 (17.6%) were attributed to differences in causal degree. In contrast, DeepDiff-SHAP detected 19 association changes, with 9 (47.3%) resolved as true causal degree differences, demonstrating a significantly higher sensitivity in uncovering meaningful causal changes (P-value = 0.02). A similar pattern emerged in the sepsis dataset: DeepDiff-SHAP resolved 1 causal degree difference out of 1 association change, whereas the original DCI resolved 8 out of 18 association changes (P-value = 0.2). Although the weaker significance in the sepsis dataset may reflect the presence of only a single nonlinear association difference, these findings nonetheless underscore the improved sensitivity of DeepDiff-SHAP, particularly for detecting complex, nonlinear causal relationships.

3.3.2. *Benchmarking DeepDiff-SHAP against CausalMGM: comparing objectives and outputs with an existing causal inference method*

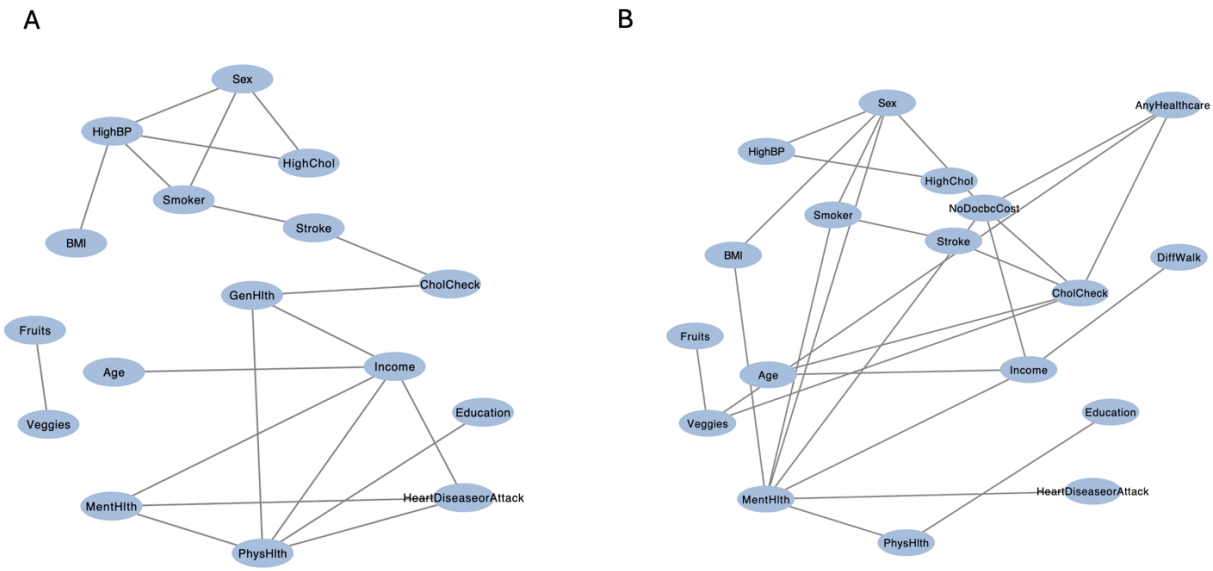

Figure 4. Differential causal graphs on CDC Diabetes data as identified by CausalMGM. (A) non-diabetic group (B) diabetic and prediabetic group.

We benchmarked DeepDiff-SHAP against CausalMGM[6] by running CausalMGM separately on the diabetic/prediabetic and non-diabetic subgroups (**Fig. 4**). CausalMGM learns which features are associated via regression and then attempts to assign a causal direction to those associations using the Peter-Clark (PC) algorithm[35]. Importantly, PC does not determine causal directions when the data does not have enough empirical evidence to do so. In our CDC Diabetes dataset, neither subgroup produced any oriented edges, suggesting that PC as part of CausalMGM has limited power to infer causal directions in this context. By extension, it would be even more difficult to compare causal differences between the two groups using CausalMGM. Despite the absence of oriented edges, CausalMGM did recover several undirected associations. Many were shared across groups and likely reflect background co-variation rather than diabetes-specific biology (e.g., high blood

pressure and high cholesterol, labeled “HighBP” and HighChol”; history of stroke and lifetime cigarette consumption, labeled “Stroke” and “Smoker”; how many days in the last 30 days were an individual’s physical and mental health considered “not good”, respectively labeled “PhysHealth” and “MentHlth”). CausalMGM also reported group-specific patterns; in the diabetic and prediabetic subgroup, associations centered on having any kind of healthcare, labeled “AnyHealthcare”, with links to CholCheck, NoDocbcCost (doctor visit skipped due to high cost), and vegetable consumption once or more per day, labeled “Veggies”, whereas in the non-diabetes subgroup, associations primarily involved general health condition, labeled “GenHlth” (e.g., connections to CholCheck, Income, PhysHlth). These findings are suggestive but remain undirected under CausalMGM. DeepDiff-SHAP advances this framework in two ways. First, DeepDiff-SHAP could determine causal directions on two diabetes-specific edges that CausalMGM only flagged as undirected: CholCheck to Age and CholCheck to AnyHealthcare. The CholCheck to Age direction is not causally plausible (a test cannot cause age) and likely reflects latent confounding or reverse causation. However, CholCheck to AnyHealthcare is consistent with a plausible behavioral pathway (i.e. lipid screening may lead to more health-care engagement). In both cases, DeepDiff-SHAP provides specified causal directions that can serve as hypotheses for relevant further investigation and work. Additionally, DeepDiff-SHAP determined the causal direction of nonlinear associations, some of which are known to be associated with diabetes[31,36]. Altogether, these results (1) validate several DeepDiff-SHAP findings against an established graphical-model baseline, (2) demonstrate greater power to orient effects where CausalMGM remains agnostic, and (3) show enhanced interpretability through subgroup-specific direction estimates and detection of nonlinear structure relevant to diabetes biology.

4. Discussion

We introduce DeepDiff-SHAP as a method that builds on regression-based DCI with a deep learning-based differential causal inference framework (**Fig. 1B**). Our method’s modular design, split into three steps with flexible parameter initialization settings, allows researchers to systematically evaluate subgroup-specific differences in feature importance and directional relationships, creating a highly customizable approach to evaluating causal data structure. Additionally, the results observed from our comparative analysis across two unique patient datasets, the CDC Diabetes Health Indicators Dataset and a UK Biobank cohort of sepsis patients stratified by hypertension, demonstrate DeepDiff-SHAP’s ability to identify biologically meaningful and study-supported causal relationships that regression-only DCI may overlook.

DeepDiff-SHAP’s performance in the analysis of the CDC Diabetes dataset offers novel insights into the differential roles of classical foundational variables such as age and sex. These variables act as upstream regulators, often influencing many downstream pathways at once. They can moderate, mediate, or interact with other risks, rather than affecting only a single chain of events. DeepDiff-SHAP models these nonlinear, interactive, and layered relationships through a unified DL-based framework. This framework helps uncover how different factors can influence multiple pathways at once, highlighting regulatory patterns that may vary across subgroups. By capturing these effects without having to specify every possible mediator or interaction in advance, it offers a practical way to study the complexity of chronic diseases like diabetes. Diabetes is a persistent metabolic disease that compounds over time; individuals with diabetes are more likely to experience a steeper deterioration in general health as they age, compared to individuals without diabetes. Additionally, aging itself brings about general decline in metabolic efficiency, immune function, and tissue repair,

all of which can be exacerbated by the presence of diabetes. By leveraging DL's multilayer modeling capacities, we reveal a synergistic effect between age and diabetes, where the impact of aging on general health is causally differential. In contrast, individuals without diabetes may experience a more gradual, less pronounced decline in general health with aging. Similarly, regarding the unique causal difference result between sepsis patients with and without hypertension, several studies have made the link between serum alkaline phosphatase (ALP) and coronary artery disease, with hypertension recognized as a well-known contributor to the development of coronary artery disease[37,38]. Additional studies in humans and mice have linked ALP with elevated cholesterol levels, particularly in those with dyslipidemia, and observe that high ALP in dyslipidemia patients leads to hypertension and coronary heart disease[39,40].

An important design consideration in DCI is deciding which variable defines the groups being compared. For the CDC diabetes dataset analysis, we looked at differences between individuals with and without diabetes, using diabetes status as the grouping variable. In this construct, the key question becomes whether the relationship between an exposure (X) and an outcome (Y) differs by diabetes status. If, instead, the question of interest centered around sex-specific differences, then the grouping variable would have to be defined as sex. Misspecification of the grouping variable with the intended modifier can lead to estimating a completely different causal relationship than originally intended and, in some cases, introduce collider bias if the outcome itself is used to define groups. For this reason, we recommend that the grouping variable of interest be specified a priori and that steps be taken to ensure balance across groups, such as through reweighting or matching.

In practice, DeepDiff-SHAP is designed to produce two outputs that are functionally relevant to clinical settings. First, it produces subgroup-specific causal networks that diagram how exposures influence outcomes differently across a chosen modifier. Second, it translates the identified subgroup-level effects into targeted risk predictions using SHAP-based effect size estimation. By structuring the analysis around the modifier of interest, but still reinforcing the defined disease as the *outcome*, this approach avoids common pitfalls like collider bias and ensures that causal results are in line with the question of interest. In conjunction, DeepDiff-SHAP's outputs support clearer communication of subgroup-specific risks and help prioritize the most important risk factors for patient care.

Despite the promise of DeepDiff-SHAP in identifying subgroup-specific differences in causal structure, our current framework has some limitations. First, like other computational approaches to causal discovery, DeepDiff-SHAP faces fundamental limits, particularly when reverse causality or feedback loops are present. Most algorithms based on directed acyclic graphs (DAGs) assume acyclicity, faithfulness, and often causal sufficiency. These assumptions prevent them from representing feedback processes and make it challenging to orient causal directions when multiple graphs are observationally indistinguishable. In practice, claiming and assigning directionality to edges typically warrants additional sources of information, such as interventions, time ordering, or invariance across environments. However, our results remain vulnerable to hidden confounding, measurement error, or selection bias. Established methods such as the PC[35] algorithm, the Fast Causal Inference (FCI)[41] algorithm, and the Greedy Equivalence Search (GES)[42,43] algorithm formalize some of this, but often lose rigor with higher dimensional data structures such as -omics data. Differential causal inference (DCI) methods that compare two conditions inherit these same identification constraints from their base graphs and can misattribute distributional shifts as causal edge differences when assumptions fail. Thus, reverse causality and hidden common causes remain crucial obstacles to solve concerning causal discovery from observational data. Second, the

computational efficiency of the SHAP-based conditional invariance testing step remains a major challenge. Specifically, estimating conditional SHAP values for each candidate variable pair across multiple conditioning subsets is computationally expensive, particularly for large-scale datasets with many variables and complex feature interactions. Third, our current approach uses an ad hoc restriction on the conditioning set to reduce complexity: we limit candidate conditioning variables to those involved in edges identified in the difference undirected graph (Δ-UG). While this helps avoid an exponential increase in the number of SHAP computations, it may miss subtler or higher-order conditional dependencies. This limitation can inherently bias results towards more prominent signal changes while underestimating other nuanced shifts in conditional structure. To address this, future iterations of DeepDiff-SHAP could benefit from dimensionality reduction techniques such as variable screening, supervised embedding, or attention-based feature selection before applying conditional SHAP. Finally, we will also investigate adaptive strategies for selecting conditioning sets that leverage measures such as mutual information or latent feature representations. The goal is to balance out computational feasibility and the capacity to capture potentially hidden dependencies. Methods such as autoencoder-based dimensionality reduction, graph neural network–derived embeddings, or Bayesian network–informed priors may help us develop principled ways to limit the conditioning set space while still retaining the most relevant differential dependency structures in data sources of interest.

As medicine increasingly moves toward personalized interventions, understanding how risk factors or biological pathways operate differently across patient subpopulations, such as those with or without comorbidities like diabetes or hypertension, is essential to avoid generalized solutions that can be ineffective or even harmful. Traditional regression-based methods average over heterogeneity, preventing subtle but important biological differences from being uncovered. Our approach, which integrates the theoretical rigor of differential causal inference with the interpretable power of deep learning and SHAP, allows researchers to discover changes in causal structure that vary with disease state, comorbidity, or population subgroup. These insights can directly inform the design of more precise diagnostic criteria, risk prediction tools, and treatment strategies, ultimately improving clinical outcomes by ensuring the appropriate interventions are delivered to the appropriate patients.

5. Acknowledgments

The CDC Diabetes Health Indicators dataset comes from the University of California, Irvine Machine Learning Repository and the CDC 2014 Annual Data report. UK Biobank data comes from the UK Biobank Resource under Application Number 83829. This research was supported in part by the University of Pittsburgh Center for Research Computing, RRID:SCR_022735, through the resources provided. Specifically, this work used the HTC cluster, which is supported by NIH award number S10OD028483.

6. Code and Data Availability

DeepDiff-SHAP code and examples can be accessed at: https://github.com/ads303/DeepDiff-SHAP. UK Biobank data is not permitted for public dispensing. CDC Diabetes survey data can be accessed at: https://archive.ics.uci.edu/dataset/891/ for immediate download.

References

1 Silander, K. *et al.* Gender Differences in Genetic Risk Profiles for Cardiovascular Disease. *PLOS ONE* **3**, e3615 (2008).

2 Peddinti, G. *et al.* Early metabolic markers identify potential targets for the prevention of type 2 diabetes. *Diabetologia* **60**, 1740-1750 (2017).

3 Sedgewick, A. J. *et al.* in *Bioinformatics* Vol. 35 1204-1212 (2019).

4 Zheng, X., Aragam, B., Ravikumar, P. K. & Xing, E. P. in *Advances in Neural Information Processing Systems* Vol. 31 (eds S Bengio *et al.*) (Curran Associates, Inc., 2018).

5 Moccia, C. *et al.* Machine learning in causal inference for epidemiology. *European Journal of Epidemiology* **39**, 1097-1108 (2024).

6 Ge, X., Raghu, V. K., Chrysanthis, P. K. & Benos, P. V. CausalMGM: an interactive web-based causal discovery tool. *Nucleic Acids Res* **48**, W597-W602 (2020).

7 Belyaeva, A., Squires, C. & Uhler, C. in *Bioinformatics* Vol. 37 3067-3069 (2021).

8 Davar, D. *et al.* Neoadjuvant vidutolimod and nivolumab in high-risk resectable melanoma: A prospective phase II trial. *Cancer Cell* **42**, 1898-1918.e1812 (2024).

9 Ruberto, V. L., Jha, M. K. & Murrough, J. W. Pharmacological Treatments for Patients with Treatment-Resistant Depression. *Pharmaceuticals* **13**, 116 (2020).

10 Olecka, M. *et al.* Nonlinear DNA methylation trajectories in aging male mice. *Nature Communications* **15**, 3074 (2024).

11 Zablotskii, V., Gorobets, O., Gorobets, S. & Polyakova, T. Effects of Static and Low-Frequency Magnetic Fields on Gene Expression. *Journal of Magnetic Resonance Imaging* (2025).

12 Cheng, W.-C. *et al.* Non-linear association between long-term air pollution exposure and risk of metabolic dysfunction-associated steatotic liver disease. *Environmental health and preventive medicine* **29**, 7-7 (2024).

13 Demateis, D., Keller, K. P., Rojas-Rueda, D., Kioumourtzoglou, M. A. & Wilson, A. Penalized distributed lag interaction model: Air pollution, birth weight, and neighborhood vulnerability. *Environmetrics* **35**, e2843 (2024).

14 Chesnaye, N. C. *et al.* Non-linear relationships in clinical research. *Nephrology Dialysis Transplantation* **40**, 244-254 (2025).

15 Zhang, Y. *et al.* Analysis of the nonlinear relationships between insulin resistance indicators such as LAP and TyG and depression, and population characteristics: a cross-sectional study. *European Journal of Medical Research* **30**, 513 (2025).

16 Oh, T. R. Integrating predictive modeling and causal inference for advancing medical science. *Childhood Kidney Diseases* **28**, 93-98 (2024).

17 Nkoy, F. L., Stone, B. L., Zhang, Y. & Luo, G. A roadmap for using causal inference and machine learning to personalize asthma medication selection. *JMIR Medical Informatics* **12**, e56572 (2024).

18 Deshpande, S., Li, Z. & Kuleshov, V. Multi-Modal Causal Inference with Deep Structural Equation Models. *arXiv preprint arXiv:2203.09672* (2022).

19 Zang, C., Wang, H., Pei, M. & Liang, W. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 19027-19036.

20 Lundberg, S. M. & Lee, S.-I. Vol. 30 (eds I. Guyon *et al.*) (2017).

21 Sundararajan, M. & Najmi, A. in *International conference on machine learning.* 9269-9278 (PMLR).

22 Belyaeva, A., Squires, C. & Uhler, C. DCI: learning causal differences between gene regulatory networks. *Bioinformatics* **37**, 3067-3069 (2021).

23 Lütkepohl, H. *New introduction to multiple time series analysis*. (New York : Springer, 2005).

24 Yuhao Wang, C. S., Anastasiya Belyaeva, Caroline Uhler. in *NeurIPS 2018* Vol. 31 (Advances in Neural information Processing Systems, Montreal, 2018).

25 Scott M Lundberg, S.-I. L. in *International Conference on Neural Information Processing Systems* Vol. 30 (Advances in Neural Information Processing Systems, Long Beach, CA, 2017).

26 Peters, J., Bühlmann, P. & Meinshausen, N. Causal Inference by using Invariant Prediction: Identification and Confidence Intervals. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **78**, 947-1012 (2016).

27 Meek, C. in *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence* 403–410 (Morgan Kaufmann Publishers Inc., Montréal, Qué, Canada, 1995).

28 Diego Colombo, M. H. M. Order-Independent Constraint-Based Causal Structure Learning. *Journal of Machine Learning Research* **15**, 3921—3962 (2014).

29 Burrows, N. R., Hora, I., Geiss, L. S., Gregg, E. W. & Albright, A. Incidence of End-Stage Renal Disease Attributed to Diabetes Among Persons with Diagnosed Diabetes - United States and Puerto Rico, 2000-2014. *MMWR Morb Mortal Wkly Rep* **66**, 1165-1170 (2017).

30 Guo, J. *et al.* Aging and aging-related diseases: from molecular mechanisms to interventions and treatments. *Signal Transduct Target Ther* **7**, 391 (2022).

31 Madonna, R. *et al.* Impact of Sex Differences and Diabetes on Coronary Atherosclerosis and Ischemic Heart Disease. *J Clin Med* **8** (2019).

32 Yoshida, Y., Chen, Z., Fonseca, V. A. & Mauvais-Jarvis, F. Sex Differences in Cardiovascular Risk Associated With Prediabetes and Undiagnosed Diabetes. *Am J Prev Med* **65**, 854-862 (2023).

33 Ballotari, P., Venturelli, F., Greci, M., Giorgi Rossi, P. & Manicardi, V. Sex Differences in the Effect of Type 2 Diabetes on Major Cardiovascular Diseases: Results from a Population-Based Study in Italy. *Int J Endocrinol* **2017**, 6039356 (2017).

34 Ahlberg, C. D. *et al.* Linking Sepsis with chronic arterial hypertension, diabetes mellitus, and socioeconomic factors in the United States: A scoping review. *Journal of Critical Care* **77**, 154324 (2023).

35 Pearl, J. (Cambridge University Press, Cambridge, 2009).

36 Tramunt, B. *et al.* Smoking and Diabetes: Sex and Gender Aspects and Their Effect on Vascular Diseases. *Can J Cardiol* **39**, 681-692 (2023).

37 Chen, Y. *et al.* Patients with comorbid coronary artery disease and hypertension: a cross-sectional study with data from the NHANES. *Ann Transl Med* **10**, 745 (2022).

38 Weber, T. *et al.* Hypertension and coronary artery disease: epidemiology, physiology, effects of treatment, and recommendations : A joint scientific statement from the Austrian Society of Cardiology and the Austrian Society of Hypertension. *Wien Klin Wochenschr* **128**, 467-479 (2016).

39 Adamidis, P. S. *et al.* Association of Alkaline Phosphatase with Cardiovascular Disease in Patients with Dyslipidemia: A 6-Year Retrospective Study. *J Cardiovasc Dev Dis* **11** (2024).

40 Bessueille, L. *et al.* Inhibition of alkaline phosphatase impairs dyslipidemia and protects mice from atherosclerosis. *Transl Res* **251**, 2-13 (2023).

41 Zhang, J. in *Artificial Intelligence* Vol. 172 1873-1896 (2008).
42 Ramsey, J., Glymour, M., Sanchez-Romero, R. & Glymour, C. in *International Journal of Data Science and Analytics* Vol. 3 121-129 (2017).
43 Ge, E., Li, Y., Wu, S., Candido, E. & Wei, X. Association of pre-existing comorbidities with mortality and disease severity among 167,500 individuals with COVID-19 in Canada: A population-based cohort study. *PLoS One* **16**, e0258154 (2021).