

# ExDBSCAN: Explaining DBSCAN with Counterfactual Reasoning

Pernille Matthews\*

matthews@cs.au.dk

Aarhus University

Department of Computer Science

Aarhus, Denmark

Lena Krieger\*

l.krieger@fz-juelich.de

IAS-8, Forschungszentrum Jülich

Jülich, Germany

LMU Munich, MCML

Munich, Germany

Tommaso Amico\*

tomam@cs.au.dk

Aarhus University

Department of Computer Science

Aarhus, Denmark

Arthur Zimek

zimek@imada.sdu.dk

University of Southern Denmark

Department of Computer Science and

Mathematics

Odense, Denmark

Thomas Seidl

seidl@dbz.ifl.lmu.de

LMU Munich, MCML

Munich, Germany

Ira Assent

ira@cs.au.dk

Aarhus University

Department of Computer Science

Aarhus, Denmark

IAS-8, Forschungszentrum Jülich

Jülich, Germany

## Abstract

Clustering is an unsupervised technique for grouping data points by similarity. While explainability methods exist for supervised machine learning, they are not directly applicable to clustering, making it challenging to understand cluster assignments. This interpretability gap is particularly evident in the popular density-based method DBSCAN, which assigns points as inliers (cluster members in dense regions) or outliers (noise points in sparse regions). DBSCAN does not provide insight into why a particular point receives its assignment or if it is robust to small changes in the data. To address the lack of explainability, we introduce ExDBSCAN, a density-aware, post-hoc explanation method. ExDBSCAN offers actionable counterfactual explanations, with theoretical guarantees for validity. It generates multiple counterfactuals using a density-connected weighted graph, adopting a physics-inspired model that repels counterfactual candidates from one another (diversity), while pulling them toward the instance to explain (proximity). Empirical evaluation on 30 tabular datasets comparing against four baselines shows that ExDBSCAN outperforms all baselines while attaining perfect validity and retrieving diverse, proximal counterfactuals.

## CCS Concepts

• **Computing methodologies** → **Cluster analysis**; • **Mathematics of computing** → **Computing most probable explanation**.

## Keywords

XAI, Counterfactuals, Explainability, Clustering, DBSCAN

## ACM Reference Format:

Pernille Matthews, Lena Krieger, Tommaso Amico, Arthur Zimek, Thomas Seidl, and Ira Assent. 2026. ExDBSCAN: Explaining DBSCAN with Counterfactual Reasoning. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2 (KDD '26)*, August 09–13, 2026, Jeju Island, Republic of Korea. ACM, New York, NY, USA, 12 pages. <https://doi.org/10.1145/3770855.3817692>

## 1 Introduction

Clustering is a cornerstone of unsupervised machine learning, enabling discovery of patterns and groupings in data without relying on predefined labels. Our work focuses on Density-Based Spatial Clustering of Applications with Noise (DBSCAN) [6], an algorithm known for its ability to find clusters of arbitrary shape and to identify outliers (also denoted as noise) in the process. DBSCAN is popular across domains ranging from computer vision [7] to bioinformatics [8] and fraud detection [20]. Nevertheless, a limitation of DBSCAN (and clustering methods generally) is their lack of explanation, i.e., the absence of information on *why* a particular point belongs to a particular cluster or not, which is crucial for interpretability, user trust, and actionable insights [23, 24].

In supervised learning, counterfactual explanations (CEs) have emerged as an impactful tool, identifying minimal changes to an input that would alter the model's prediction, offering intuitive "what-if" scenarios [9, 12, 26, 29]. Counterfactual reasoning builds user trust in the model and reflects how human stakeholders formulate explanations, making counterfactuals easy to interpret [23]. For classification tasks, a typical CE shows how a rejected loan application could be altered to become approved.

CEs are beneficial for clustering. Beyond explaining individual points, multiple counterfactuals can reveal alternative ways in which a point could transition between clusters, offering complementary perspectives on the underlying structure. Furthermore, analysing how counterfactuals change across several explained points can suggest which features are most relevant to cluster membership. These patterns not only enhance interpretability but also have the potential to help practitioners make more informed decisions about data preprocessing, feature engineering, and outlier treatment. While a formal user study is beyond the scope of this

\* Authors contributed equally to this research.



This work is licensed under a Creative Commons Attribution 4.0 International License. *KDD '26, Jeju Island, Republic of Korea*

© 2026 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-2259-2/2026/08

<https://doi.org/10.1145/3770855.3817692>

work, we illustrate this interpretive potential through concrete counterfactual examples on tabular datasets.

Although DBSCAN provides transparent mechanisms including interpretable rules, they do not answer the fundamental question of *what should I change*. For noise points specifically, DBSCAN identifies isolation but cannot indicate which features lead to isolation or how to resolve it.

Devising CEs for DBSCAN presents major challenges. First, the cluster assignment process is not based on a differentiable loss function, preventing the counterfactual method from having access to model gradients, often used to guide counterfactual search [26, 36]. Moreover, while classification algorithms often produce class probabilities, a membership probability, used by several model-agnostic counterfactual methods [26, 38], is difficult to define in hard clustering contexts. Furthermore, DBSCAN can label isolated points as noise, which may be scattered across the whole feature space and share few similarities, giving them a boundary very different from classes in a classification scenario.

Model-agnostic Bayesian and genetic counterfactual methods for classification generally do not need access to gradients or class probabilities. Their application, however, is mostly computationally expensive [30]. Furthermore, different clustering algorithms rely on different concepts of similarity. This can impact counterfactual search, for example when ensuring diversity, i.e., making sure that different counterfactuals for the same input are not repetitive but instead add meaningful actionable alternatives. In the DBSCAN case, two samples could be close in the traditional Euclidean distance leveraged by most model-agnostic methods, while at the same time, being connected only by a long density-connected path, making the two points effectively distant.

In this work, we address these challenges by leveraging counterfactual reasoning to explain density-based clustering. We propose **Explaining DBSCAN** with counterfactual reasoning (ExDBSCAN), the first method tailored to generate CEs for DBSCAN. Given a specific point, DBSCAN’s assignments, and its parameter settings, ExDBSCAN generates counterfactuals that are both proximal (close to the original instance for realistic changes) and diverse (providing multiple non-redundant alternatives), while accounting for DBSCAN’s similarity definition. To achieve this, ExDBSCAN models the problem as a physical optimisation system, defining an undirected weighted graph that captures DBSCAN’s density-based similarity structure. The system treats candidate counterfactuals as charged-like particles that repel each other to ensure diversity, and uses spring-like forces to maintain proximity to the explained instance. Our approach provably provides perfect validity, i.e., every counterfactual attains the correct cluster assignment, for both noise-to-cluster and cluster-to-cluster transitions. Furthermore, ExDBSCAN respects feature constraints by handling non-actionable attributes, e.g., the number of previous hospital visits in a tabular dataset of clinical analysis, where changing these attributes would make the explanation unrealistic, with the goal of ensuring explanations remain practical and applicable.

Our contributions include:

- We introduce ExDBSCAN, the first method to generate counterfactual explanations designed for DBSCAN, covering both noise-to-cluster and cluster-to-cluster transitions.

- ExDBSCAN generates diverse counterfactuals under our novel physics-inspired model that repels candidate counterfactuals from one another (diversity) while pulling them toward the point to explain (proximity), while also accounting for cluster structure captured in our graph representation.
- On 30 OpenML tabular datasets, ExDBSCAN outperforms all baselines by obtaining lower proximity, higher multi-counterfactual diversity and perfect validity.

## 2 Related Work

**Supervised Counterfactual Explanations.** In supervised learning, Wachter et al. [36] formally introduce CEs to answer “what is the smallest change to input  $x$  that would alter the model’s prediction?”, and subsequent work optimises criteria such as *proximity*, *validity*, *diversity*, and *actionability* [9, 22, 39]. The method by Wachter et al. [36] utilises gradients to guide the counterfactual search and thereby only applicable to differentiable methods.

DiCE [26], extends these concepts to account for *diversity*, returns a set of diverse CEs. While DiCE was originally designed for differentiable models only, it has since been adapted to be model-agnostic. FACE [29] follows data-supported feasible paths on the empirical manifold and returns a plausible path together with a counterfactual endpoint that alters model prediction. MACE [38] uses a reinforcement-learning policy and returns one or more counterfactuals with minimal changes. The model-agnostic Growing Spheres algorithm [17] explores the input space by expanding hyperspheres until the model decision boundary is reached.

Another model-agnostic counterfactual method is BayCon [30]. Unlike gradient-based methods, BayCon uses Bayesian optimisation with probabilistic feature sampling to explore the feature space without requiring differentiable model outputs. It treats the clustering algorithm as a black box, repeatedly querying it with candidate counterfactuals and using the discrete cluster assignments to guide its search. However, it does not exploit DBSCAN’s specific density-connectivity structure that ExDBSCAN leverages for efficient and theoretically grounded counterfactual generation.

All methods need to repeatedly query a supervised model for class probabilities at candidate points and then choose based on closeness to the target class. By contrast, DBSCAN provides only discrete, connectivity-based memberships (core, border, noise) defined by  $\epsilon$  and *minPts*, with no continuous quantity to optimise against. A workaround could be to fit a surrogate classifier to predict cluster memberships and then apply supervised CE methods; however, the surrogate’s decision regions would then reflect its own similarity notion rather than density-reachability.

**Explainable Clustering.** The need to understand and explain clustering results has motivated various approaches to clustering interpretability. Unlike supervised learning, where explanations often focus on decision boundaries, clustering explanations must describe why groups form and what distinguishes them.

Global explainable clustering methods aim to summarise entire clusterings or cluster characteristics in human-understandable forms. Moshkovitz et al. [25] replace *k*-means/*k*-medians clusters with compact interpretable decision trees. Similarly, Bandyapadhyay et al. [4] search for small trees that match a given partitioning.

Prototype-based methods characterise each cluster by representative examples or “criticism” points, e.g., Kim et al. [14] select prototypical instances and outliers for each cluster. Such global explanations convey what defines clusters overall, but they do not explain individual assignments.

In contrast, local explainability focuses on specific data points. Agnostic attribution methods like FACT [31] quantify feature influence on a point’s cluster assignment by perturbing features and observing changes in the clustering. Interactive tools like Cluster-Vision [16] visualise effects of input parameters on clusters and outliers.

While global summaries, prototypes, feature attributions, and visual analytics improve understanding of clustering results, they do not provide actionable local changes for a given point. No existing method can explain how a particular DBSCAN inlier or outlier could change assignment through minimal feature changes, with guarantees of valid membership under DBSCAN’s density-reachability. This gap motivates ExDBSCAN.

**Counterfactuals for Clustering or Outlier Detection.** Some recent work studies CEs for clustering. For prototype/mixture-style clustering, Vardakas et al. [35] provide the first analytic formulations: for  $k$ -means, they derive a closed-form move that crosses the Voronoi boundary into a target cluster, and for Gaussian clustering, devise and solve a non-linear formulation.

These methods leverage geometry and thus do not apply to density-based algorithms, where membership is defined by density-reachability rather than distances to prototypes.

Spagnol et al. [32] build on BayCon and propose model-specific scores: a distance-based score for  $k$ -means (scaled distance to the target centroid) and a membership-probability score for HDBSCAN\* (using the algorithm’s soft-clustering outputs). These scores guide a zero-order Bayesian optimisation loop to find CEs. While this makes BayCon-style search applicable to HDBSCAN\*, it uses HDBSCAN\*’s probabilistic model, unavailable for DBSCAN. Moreover, it provides no density-reachability guarantees and does not explain DBSCAN’s core/border/noise status.

Beyond clustering, counterfactuals have been explored for unsupervised outlier detection. Zhou et al. [40] propose EACE which for a given outlier generates one or more counterfactual samples that an anomaly detection model would classify as an inlier. Angiulli et al. [3] similarly “repair” an outlier via a subset of features and adjusted values (a “mask”) that would turn it into an inlier. Amico et al. [2] propose counterfactuals for LOF using the density-based properties of the method.

Both EACE and the density-contrastive “repair” methods target outlier detection models that expose a continuous anomaly score (e.g., one-class SVM, kernel density estimators, autoencoders, LOF-style scores). Their goal is to decrease the anomaly score so that a point crosses a detector-specific decision threshold and becomes “normal,” suitable for detectors where “inlierness” is defined by that score. DBSCAN does not provide such a score, soft memberships, or gradients; it yields discrete core/border/noise assignments based on  $\epsilon$ -reachability and  $minPts$ . Consequently, these counterfactual outlier detection methods optimise an objective that is orthogonal to ours, since they can make a point “less anomalous” in a detector’s sense without guaranteeing that it becomes density-connected to a

DBSCAN cluster. In contrast, ExDBSCAN includes both cluster-to-cluster and noise-to-cluster transitions.

### 3 Counterfactual Explanations for DBSCAN

We describe DBSCAN and counterfactual explanations (Sect. 3.1), before introducing density-based counterfactuals (Sect. 3.2) and our novel method, ExDBSCAN (Sect. 3.3).

#### 3.1 Preliminaries

Consider dataset  $X = (x_1, \dots, x_l), x_i \in \mathbb{R}^n$  and cluster assignment from points  $C : X \rightarrow \{-1, 0, \dots, m-1\}$  to one of  $m$  cluster labels, or to noise ( $-1$ ). The popular density-based clustering approach DBSCAN [6] groups points based on a similarity notion modelled as density-reachability. Clusters are maximal areas of density-connected points, separated from other clusters through sparse areas. It uses two parameters:  $\epsilon$ , which defines the radius of local density assessment, and  $minPts$ , which specifies the minimum number of points required to form a dense region. The  $\epsilon$ -neighbourhood are the points within radius  $\epsilon$  of a point  $p$ :  $\mathcal{N}_\epsilon(p) = \{q \mid \|p - q\|_2 \leq \epsilon, \text{ with } p, q \in X\}$ . DBSCAN distinguishes between dense *core points*, *border points* close to core points, and *noise points* or outliers, that do not belong to any cluster:

**Definition 3.1** (Core, border, and noise points). The set of *core points*  $Q$  are points  $p$  that have at least  $minPts$  points in their  $\epsilon$ -neighbourhood:  $Q = \{p \in X \mid |\mathcal{N}_\epsilon(p)| \geq minPts\}$ .

Points in  $X \setminus Q$  are *border points*  $B$ , if they have at least one core point in their neighbourhood, else they are *noise*.

*Counterfactual Explanations (CEs).* CEs were first proposed as an optimisation problem by Wachter et al. [36] for classifiers:

**Definition 3.2** (Counterfactual Explanations). A set of  $k$  CEs for a point  $p$  under a classification model  $f$ , denoted as  $C_k(p)$  minimises a cost function for which the classification outcome changes:

$$C_k(p) = \underset{p'_1, \dots, p'_k}{\operatorname{argmin}} \operatorname{cost}(p, p') \text{ subject to } f(p) \neq f(p') \quad (1)$$

Counterfactual generation methods vary in the choice of cost function, typically balancing between desired properties, e.g., the distance between the original point  $p$  and the counterfactual  $p'$  or the pairwise distance between counterfactuals. In the context of density-based clustering, the cost function in Definition 3.2 takes into account the clustering algorithm’s concept of similarity.

#### 3.2 Characteristics of Counterfactuals

Counterfactual explanations provide value for clustering tasks by addressing the interpretability challenges in unsupervised learning. Unlike supervised learning where model decisions can be evaluated against known labels, clustering methods function without predefined classes. This leaves the task of understanding why certain groups are formed or why points are assigned to a particular cluster. CEs bridge this gap by revealing the minimal changes required to alter a point’s cluster assignment, thus explaining group differences and features that may drive cluster membership. Density-based clustering finds clusters of arbitrary shapes and can identify noise points. These capabilities can make the resulting clusterings difficult to interpret because the complex, non-convex cluster boundaries

and the distinction between core, border, and noise points are not immediately obvious. Users cannot easily identify which specific features or density thresholds have led to the assignments. Noise-to-cluster and cluster-to-cluster CEs reveal key features that determine density-based cluster membership.

**Cluster Assignment.** The counterfactual generation process requires assigning out-of-dataset points such that we can test whether a candidate lies in the target cluster. Contrary to classifiers, DBSCAN lacks a native assignment function for new points. To ensure that we explain a given clustering without inadvertently manipulating it, and avoid re-clustering, we keep the clustering fixed and define  $C$ : point  $p$  belongs to cluster  $i$  iff there exists a core point  $q$  with  $C(q) = i$  within the  $\epsilon$ -neighbourhood, such that  $q \in \mathcal{N}_\epsilon(p)$ .

We restrict membership to core-point neighbourhoods, as placing a counterfactual near a border does not ensure density-connectivity, and may violate DBSCAN's membership definition. Keeping a fixed clustering structure also avoids the computational and conceptual pitfalls of re-clustering (e.g., merges or new clusters), ensuring that we explain the original structure rather than one altered by the CEs.

Generating multiple diverse CEs, rather than a single instance, enhances explanatory value by probing cluster-assignment boundaries from different directions, giving stakeholders a thorough view of the clustering algorithm's behaviour. Multiple examples help surface explanatory trends (e.g., features repeatedly changed likely matter for assignment) and reduce perceived arbitrariness, improving trust [23].

Regarding actionability, i.e., how feasible the proposed counterfactual change is, different paths offer different feasibilities, which largely depend on the application scenario. For example in a medical setting, in the hypothetical scenario of changing the respiratory rate versus the oxygen saturation may be more or less realistic for a given patient. Offering several options increases the likelihood that at least one constitutes a realistic, acceptable intervention, enabling feasible and actionable paths to change the outcome of the model [33].

It is thus crucial that CEs demonstrate diversity [17]. CEs that differ only marginally are redundant and fail to highlight differences among clusters [26]. High redundancy adds no valuable information, for example, two counterfactuals that vary by only 0.1 breaths per minute in respiratory rate are effectively indistinguishable and convey the same information.

While diversity is essential, CEs that stray too far from the explained point risk suggesting unrealistic changes. If a CE bears little in common to the point being explained, human stakeholders may struggle to relate the two and reason about their differences. Proximity therefore remains fundamental to counterfactual reasoning, as it increases both the realism and feasibility while reducing cognitive load for stakeholders, as there are fewer overall changes [23].

### 3.3 ExDBSCAN Counterfactuals

ExDBSCAN addresses these requirements by generating valid sets of counterfactuals that balance proximity and diversity. Users can specify their number and the target cluster to meet their needs.

**Inter-Cluster Concept of Distance.** In CE literature, density is used as a proxy for the ease of moving between points or feasibility [11, 29, 37, 40]. In DBSCAN, cluster members are density-connected, so a counterfactual that moves within a cluster stays in dense regions with similar neighbours, enabling realistic paths. However, because DBSCAN discovers non-convex shapes, two points can be close in Euclidean space yet require a long density-connected path; straight-line distance can therefore overstate similarity. Throughout this paper we use Euclidean distance as it is the most popular metric; a discussion of alternative metrics and evaluation choices are in App. 3.3.

CEs that look redundant under Euclidean distance, i.e., proposals close in straight-line distance, may still be meaningfully diverse once DBSCAN's density structure is considered. Euclidean close points can lie in different density-reachable regions and thus represent distinct alternatives. Relying on Euclidean distance risks "cutting through" clusters and ignoring connectivity, yielding explanations that misstate feasibility.

**Choice of Distance Metrics** By design, ExDBSCAN is agnostic to metrics. That being said, if DBSCAN is fit on a metric different to Euclidean, e.g., Manhattan, Cosine, Mahalanobis, or mixed-type metrics, ExDBSCAN can be easily adjusted to this. Instead of using Euclidean distance, we can employ any other distance metric to determine reference core-points as well as solving the energy-minimisation to produce diverse and proximal counterfactuals. We strongly advise using the same metric in ExDBSCAN as the one selected for DBSCAN clustering, to maintain consistency in the concepts of density-reachability and graph connectivity. Furthermore, we advise employing this chosen metric to evaluate proximity and diversity, as the underlying structure of the data and clustering depends on it.

**Reference Core Point.** Given our definition of the assignment function, a CE is assigned to the target cluster iff it falls within the  $\epsilon$ -neighbourhood of a cluster's core points. When selecting a point in the space as a CE, we naturally associate it with its nearest core point. To properly leverage the weighted graph  $G$  that captures the cluster's structure, we identify the most appropriate reference core point for each proposed CE and generate the CE example within its  $\epsilon$ -neighbourhood. Hence, a set of ExDBSCAN CEs corresponds directly to a set of vertices in graph  $G$ , namely the cluster's core points. We denote the set of CEs for point  $p$  as  $C_k(p)$ , while  $C'_k(p)$  represents the corresponding set of reference core points.

**Proximity Through Attraction.** By definition, the most proximal CEs are the ones closest to the point to explain  $p$ . Given  $k$  CEs, the set of core points  $C'(p)$  maximising proximity will thus be composed of the  $k$  nearest core points of  $p$ . As  $p$  and candidate core points are not density connected, the distance metric used is the one employed for fitting DBSCAN.

**Diversity Through Repulsion.** In CEs, a set is diverse when its members are mutually dissimilar [26]. One class of methods enforces this by maximising pairwise distances; variants of the maximum diversity problem (MDP) [17, 19]. Another models diversity via repulsion, minimising similarity between selected instances; determinantal point processes are common realisations, probabilistically favouring diverse subsets [1, 15, 26].

We observe that simply adopting (Euclidean) distance falls short for DBSCAN: its very notion of similarity is density-connectivity.

Thus, the challenge is to model proximity and diversity using density-connectivity. We adopt a repulsion-based approach that prevents redundant counterfactuals, as repulsion forces increase substantially as points converge. In contrast, maximising pairwise distances remains more forgiving toward closely positioned instances, provided that other point pairs maintain sufficient separation to preserve overall diversity. Also, repulsion interactions naturally produce more evenly distributed selections across the solution space. In our density-based setting, we implement this diversity measure using the weighted shortest path distance defined by graph  $G$  to evaluate pairwise distances in a cluster. This ensures that our diversity metric respects the cluster’s density-connectivity structure rather than relying on simple Euclidean distances that could cut through the cluster and misrepresent the actual relationships between points.

**Balancing Proximity and Diversity.** ExDBSCAN generates multiple CEs that ensure both proximity and diversity; an inherent trade-off since closer alternatives are often less diverse. We resolve this by modelling a physical system and selecting its equilibrium (the minimum-energy configuration), where the proximity-diversity balance is optimal.

For diversity, we treat candidate core points as like-charged particles with Coulomb’s electrostatic law [10], pushing already selected cores apart so the minimum-energy set is maximally spread; making the minimum energy configuration the most diverse.

For proximity, we bias chosen core points to be close to the original point, by connecting the latter with a spring to candidate core points. The further the candidate CE, the stronger the force pulling it closer to the original instance, growing linearly with the weighted path distance according to Hooke’s law [10]. Prop. 1 states how the minimum energy configuration of spring-like interactions selects the closest core points, accounting for proximity.

The aim is to find the system’s stable configuration, i.e., the one that minimises the energy of the physical system. Spring-like forces bias the system to select points closest to the original instance as springs push towards their rest state; i.e., when explained point and candidate core point have a null distance. Electrostatic-like repulsion favours core points to be pairwise distant (according to the shortest weighted path distance), as same-charged particles push each other away. The minimum energy configuration balances the springs’ desire to go back to their rest configuration with the charges’ desire to spread as much as possible. Hence, given a set  $C'(p)$  of chosen core points, and point to explain  $p$ , the energy configuration of the system has the form:

$$E_{C'} = \sum_{V_i \in C'} \sum_{V_j \in C', j > i} \frac{1}{\mathcal{D}(V_i, V_j)} + \sum_{V_i \in C'} d^2(p, V_i) \quad (2)$$

$V_i$  is the  $i$ th vertex of graph  $G$ ,  $\mathcal{D}(V_i, V_j)$  denotes the weighted shortest path distance between vertices  $V_i$  and  $V_j$ . Lastly,  $d(p, V_i)$  is the Euclidean distance between the point to explain  $p$  and vertex  $V_i$ . As these two distances may operate at different scales, with  $d(p, V_i)$  potentially larger since it connects points outside a cluster, we scale  $\mathcal{D}(V_i, V_j)$  by  $d_c/\bar{\mathcal{D}}$ , where  $d_c$  is the distance from  $p$  to the closest core point and  $\bar{\mathcal{D}}$  is the average graph weight, preventing the spring-like term from dominating. We evaluate physical constants

characterising spring and electrostatic terms in Eq. 2 as 1, as we are only interested in how energy terms scale with distance.

To find the optimal configuration of the system and thus, the set of  $C$ , we find the subset  $C'$  of vertices  $V$  minimising system energy:

$$C'_k(p) = \underset{B \in V : |B|=k}{\operatorname{argmin}} E_B \quad (3)$$

The energy term in Eq. 2 can be seen as ExDBSCAN’s cost function in the formalism introduced in Eq. 1. Fig. 1 shows how a solution of Eq. 3 can be visualised on a toy dataset where counterfactuals are produced for a noise point. On the left of Fig. 1, considering the electrostatic term only leads to maximally diverse counterfactuals. On the right, using just spring-like terms make counterfactuals maximally proximal. In the middle, balancing repulsion and attraction considering the cluster’s density structure, prompts proximal and spread-out counterfactuals. We observe the following:

**Proposition 1.** *Aligning with our desiderata on proximity, if considering only the spring-like term in Eq. 2, solving the optimisation problem in Eq. 3 is equivalent to selecting the  $k$  nearest neighbours of the point to explain  $p$ .*

**PROOF.** From Eq. 2, and by induction, when the set of core points is empty, the core minimising the elastic energy of the system is the closest to the point to explain  $p$  (minimum possible distance). As for the induction step, given a set of  $k' < k$  selected core points, the next core point to be selected, i.e., the one minimising the total elastic energy, is given by the closest non-selected core-point to  $p$ .  $\square$

Solving the optimisation problem in Eq. 3 is computationally challenging.

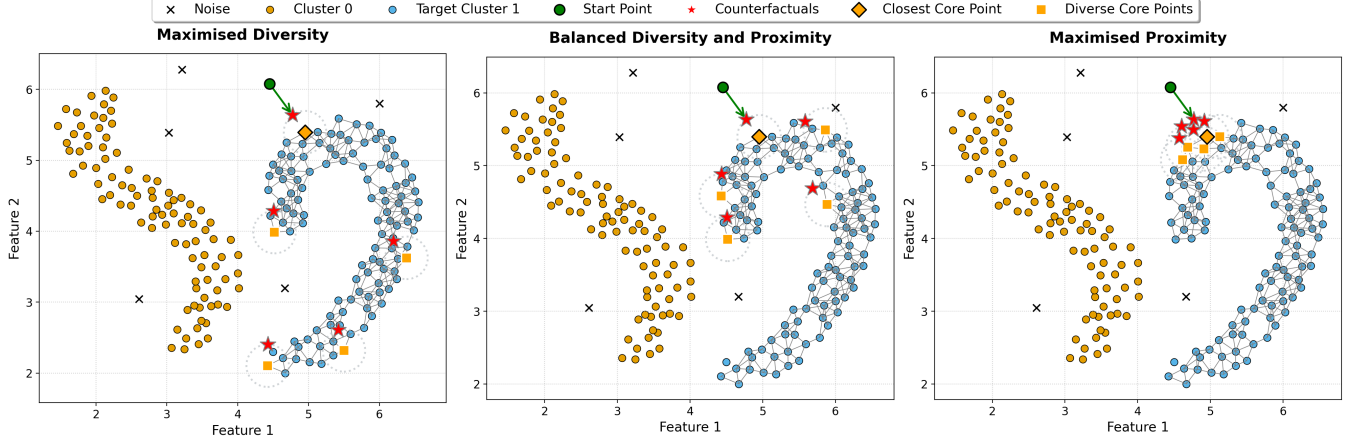
**Proposition 2.** *Solving Eq. 3, i.e., the energy minimisation problem with both attraction and repulsion, is an NP-hard problem.*

**PROOF.** (Sketch, full in Appendix A.3)

Maximum diversity problems are NP-hard [28]. Eq. 3 restricted to the electrostatic term can be mapped to the MDP. Adding springs does not change NP-hardness: If it did, the MDP could be solved in polynomial time sending spring constants to zero.  $\square$

We approximate the solution using the following greedy algorithm. Given an incomplete set of selected core points  $C'_k$  with  $k' < k$ , the next cluster point is chosen as the one minimising the total energy of the system according to Eq. 2. When choosing the instance initialising the set of core points,  $C'$  is still empty and thus no repulsive electrostatic-like force influences the choice.  $C'$  will thus be initialised with the core point closest to the point to explain since it is the one greedily minimising the energy configuration, composed by a single spring-like term. Note in the case with only one counterfactual, the greedy optimisation corresponds to the optimal one according to Eq. 2. Full pseudocode, a complexity analysis, and an empirical assessment of approximation quality are provided in App. A.2, App. A.6, and App. A.4, respectively.

**Choosing the set of counterfactuals  $C$ .** Our chosen set of core points  $C'$  leverages the clusters’ density structure to optimally balance proximity and diversity through the weighted undirected graph  $G$ . Given  $C'$ , we need to select where in each core point’s  $\varepsilon$ -neighbourhood the counterfactual is going to be placed, practically



**Figure 1: Optimising diversity (left), balanced (middle), optimising proximity (right). Cluster points in yellow and blue, resp.; noise marked with X; green point to be explained; red stars counterfactuals.**

mapping  $C'$  to the final set of counterfactuals  $C$ . Given a single core point  $q$  in  $C'$ , to further improve proximity and respecting our assignment function definition, we position the counterfactual  $p' \in C$ ,  $\varepsilon$  away in the direction of the original point  $p$ :

$$p' = p + (d_{pq} - \varepsilon)(q - p)/d_{pq}. \quad (4)$$

The counterfactual point  $p'$  is considered part of the cluster containing reference core point  $q$ , as it satisfies the density-based clustering criteria for the target cluster, effectively ensuring validity.

**Theorem 3.1.** *Every counterfactual point  $p'$  with reference core point  $q$ , such that  $C(q) = i$ ,  $i \neq -1$ , that is generated for a point  $p \in X : C(p) \neq i$  toward cluster  $i$  with the proposed method, is considered part of cluster  $i$ , implying that  $C(p') \neq C(p)$  and  $C(p') = i$ , and thus valid counterfactual.*

**PROOF.** (Sketch) The definition in Section 3.2 denotes that  $p'$  is in the same cluster as its reference core point  $q$  with  $C(q) = i$ , therefore  $C(p') \neq C(p)$  and specifically  $C(p') = i$ . ■

App. A.5 details how to retrieve counterfactuals  $C$  when some features are not actionable, and shows how ExDBSCAN creates a graph only considering reference core points whose  $\varepsilon$  neighbourhood is reachable without changing non-actionable features. Discussion on how to implement feature constraints such as monotonicity and positive correlation can be found in the extended material [21].

## 4 Experiments

**Experimental Setup.** In this paper we evaluate counterfactual explanations on 30 tabular datasets from OpenML [34]<sup>1</sup>. We further provide in the extended version [21] two additional evaluations specifically on spatial datasets [21]<sup>2</sup>. For each dataset, DBSCAN hyperparameters ( $\varepsilon$ , minPts) are selected through a grid search to maximise the DBCV clustering validity index [27].

<sup>1</sup>Dataset details and reproducible code: <https://github.com/tommasoamico/ExDBSCAN>.

<sup>2</sup>datasets, where data points represent geographical locations, are often clustered with DBSCAN as it can successfully capture their density structure.

**Baseline Methods.** ExDBSCAN is compared against seven baseline methods and variants for counterfactual generation:

- (1) **BayCon** [30], a model-agnostic Bayesian counterfactual generator. BayCon uses Bayesian optimisation to propose candidate points; to model the objective posterior, it fits a Random Forest surrogate internally.
- (2) **DiCE** [26] uses an optimisation-based counterfactual formulation and incorporates diversity into the objective. DiCE is evaluated in two variants: *DiCE-Direct*, where DBSCAN's discrete assignment is queried during optimisation, and *DiCE-Surrogate*, where DiCE optimises against a fitted surrogate classifier to obtain a smoother search space.
- (3) **Growing Spheres (GS)** [17] is a greedy method that samples instances on hyperspheres of increasing radius until a point crossing the decision boundary is found. We evaluate two variants, *GS-Direct* on DBSCAN's discrete assignments and *GS-Surrogate* on top of a fitted surrogate model.
- (4) **ExDBSCAN Random** is a heuristic baseline which samples  $k$  reference core points uniformly at random from the target cluster's core set, and constructs counterfactuals from these cores using the same core-point-based construction as ExDBSCAN (Section 3). This isolates the contribution of ExDBSCAN's structured core-point selection (proximity-diversity optimisation) from the final counterfactual construction step. We perform a deeper ablation study beyond ExDBSCAN Random which inspects how ExDBSCAN's performances change without repulsive forces and attraction forces respectively in the extended material [21].

**Surrogate Models and Training.** DBSCAN produces discrete assignments (inliers/outliers) and exposes neither class probabilities nor usable gradients. As a consequence, supervised counterfactual methods like DiCE and Growing Spheres do not have a smooth objective to optimise when applied directly; leading to unstable optimisation and difficulty crossing the clustering boundary. For fair comparison against these methods, evaluation occurs through

a surrogate setting that fits a multi-class classifier that predicts DBSCAN cluster labels and generates counterfactuals with respect to the surrogate’s probability output. This is a standard workaround of treating cluster labels as *pseudo-classes* [13, 18]. While acknowledging that the resulting decision regions may reflect the surrogate’s inductive bias rather than DBSCAN’s underlying density-reachability structure, inclusion of surrogates is necessary due to the lack of methods able to adequately compute valid DBSCAN counterfactuals. For surrogate-based variants, DBSCAN cluster labels are used as training targets; except for BayCon, which natively fits a Random Forest surrogate. We select the surrogate via grid search over feed-forward neural networks, Random Forests, and XGBoost classifiers [5], choosing the best-performing model per dataset.

**Sampling Strategy.** For each method, we use the following strategy to select the samples that will be explained. For each cluster  $C_i$  or noise partition  $N$ , we sample 10 points randomly. Each sampled point  $p \in C_i$ , targets every other cluster  $C_j$  with  $j \neq i$  yielding  $10 \times \sum_{j=1}^m (m-1) + 10 \times m$  queries per dataset. Each query asks to produce a set of  $k = 10$  CEs in the specified target cluster. The only method that cannot output a precise number of counterfactuals is BayCon, which inherently produces a large set of explanations. Thus, we evaluate BayCon by taking both 10 random counterfactuals and the 10 closest to the original instance. We refer to the two cases with *BayCon Random* and *BayCon Closest* respectively.

**Counterfactual Evaluation Metrics.** To evaluate the quality of counterfactuals across methods we utilise three common counterfactual evaluation metrics: *Validity*, *Proximity* and *Diversity*.

- (1) **Validity ( $\uparrow$  is better):** Proportion of returned  $k$  counterfactuals that reach the target cluster (i.e., the CE is part of the target cluster). When no CE is found, validity is zero for that query. Higher values are better for validity, as the goal is to obtain as many valid CEs as possible.
- (2) **Average Proximity ( $\downarrow$  is better):** Mean distance from original point  $p$  to each valid returned CE, measured in DBSCAN’s clustering feature space. The distance is the same as the one used to fit DBSCAN, here Euclidean distance. Lower values indicate that the CE more closely resembles the explained point.
- (3) **Average Diversity ( $\uparrow$  is better):** We want  $k$  CEs to offer diverse solutions in the target cluster, not minor variants of the same information. To properly evaluate diversity, we use the diversity metric based on the determinantal point process used in Mothilal et al. [26] and argued for by Kulesza and Taskar [15]. Specifically, we compute a kernel matrix  $K_{ij} = 1/(1 + d(cf_i, cf_j))$  and use its determinant as a measure of diversity. Higher determinant values indicate greater diversity among proposed counterfactuals. The more diverse the vectors inspected, the higher the volume computed by the determinant to estimate diversity.

To measure diversity, we apply a Determinantal Point Process (DPP) kernel based on the weighted shortest-path distance within the cluster graph. For points that are density-connected, like our counterfactuals, graph distance effectively captures the similarity structure of DBSCAN, whereas Euclidean distance may not accurately represent density-connectivity. Furthermore, between the

explained point and its counterfactuals there is no density connectivity, making Euclidean distance, the metric used for fitting DBSCAN, the most appropriate measure. To extend the discussion, we illustrate concretely how Euclidean distance can misrepresent diversity, where counterfactuals at opposite ends of an arc appear close in Euclidean space yet are density-distant; see extended material [21].

To realise this, for each cluster we build an undirected weighted graph  $G(V, E)$  whose vertices  $V$  are core points; edges connect two vertices if they are directly density-reachable (distance  $< \epsilon$ ), with their distance as edge weight. The distance between any two points is the weighted shortest-path in  $G$ . Thus, instances are “close” only if an easy path connects them, naturally respecting the cluster’s density structure.

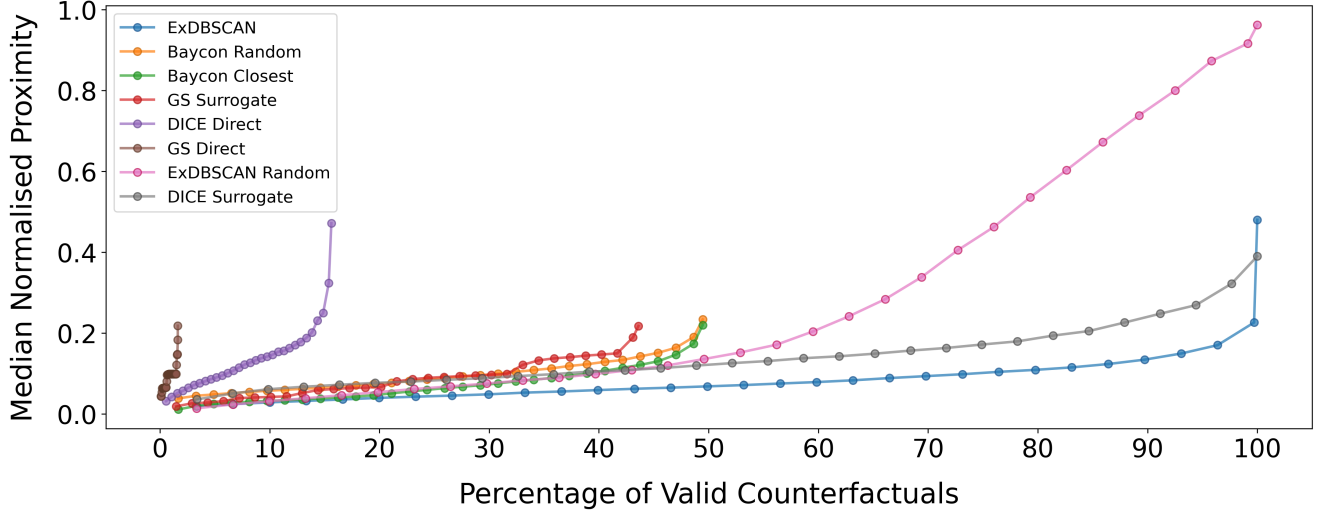
We report counterfactual quality using empirical percentile curves, which visualise the distribution of each metric across queries while simultaneously reflecting method failure rates. For each method, per-query values are sorted and plotted with the metric value on the  $y$ -axis and the cumulative percentage of all queries on the  $x$ -axis, where the  $x$ -position of the  $x$ -axis is given by  $100 \cdot r/N$  with  $N$  denoting the total number of queries. Queries for which a method fails to return a valid counterfactual, yield *undefined* metric values and do not contribute points to the curve. Consequently, curves that extend further to the right indicate higher percentage of validity i.e., more successful queries. Due to space, experiments for non-actionable scenarios can be found in Appendix A.5.4. A qualitative example on the California Housing dataset, sparsity and plausibility can instead be found in our extended material [21].

## 4.1 Validity and Proximity Evaluation

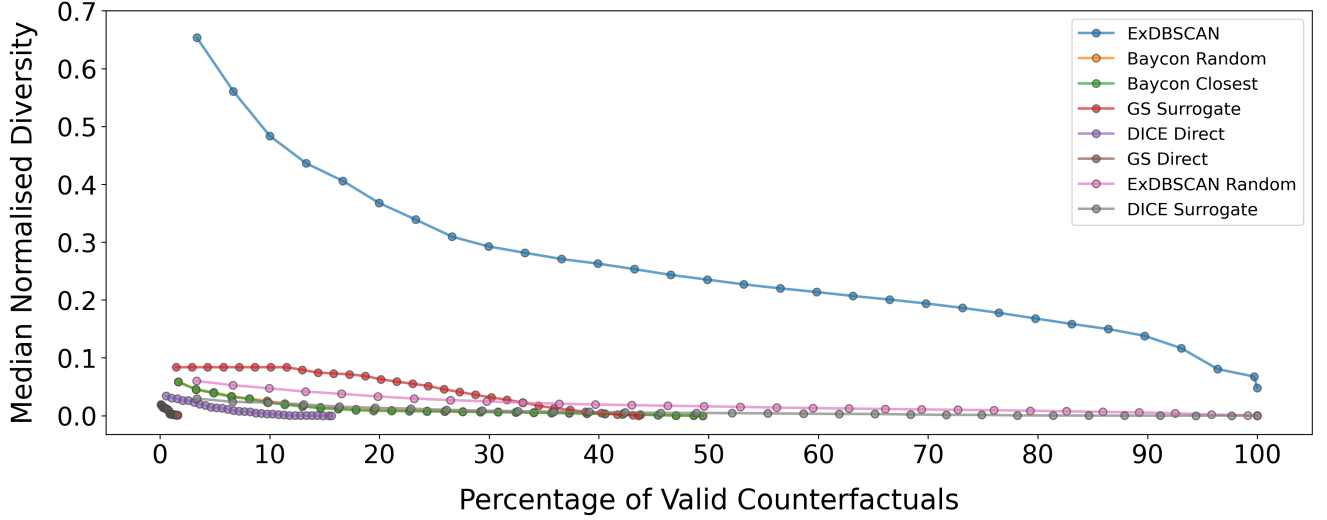
Fig. 2 reports results for *validity* and *proximity*. The  $y$ -axis indicates the *proximity* metric values, while the  $x$ -axis reports cumulative percentages of queries for which a valid counterfactual has been obtained. Lower proximity values correspond to less changes from the original instance and thus better counterfactual quality.

**Validity Performance.** ExDBSCAN achieves perfect validity (100%) across all queries, successfully generating valid counterfactuals in every case. This is guaranteed by our theoretical analysis (Theorem 3.1), which ensures that every counterfactual obtained by Eq. 4 lies within  $\epsilon$  of a core point in the target cluster and thus satisfies DBSCAN’s density-connectivity requirements. *ExDBSCAN Random* and *DiCE surrogate* also achieve 100% validity. The full table of validity values are presented in the Appendix in Table 1 in the extended version [21].

In contrast, the surrogate-based methods, *DiCE-Surrogate* and *GS-Surrogate*, as well as *BayCon* struggle severely with validity, mostly below 50% successful counterfactual generation. This performance gap originates from the fundamental mismatch of these methods optimising against continuous decision boundaries learned by surrogate classifiers or probabilistic models, which do not faithfully capture DBSCAN’s discrete, density-connectivity-based membership requirement. The surrogate’s smooth decision regions may unintentionally exclude valid density-connected points or include invalid ones, which leads to generated counterfactuals that fail on DBSCAN’s actual assignment function. *DiCE-Direct* performs poorly due to the problem not being characterized by gradients or



**Figure 2: Validity/proximity.** Cumulative percentage of successful queries (x-axis) vs. proximity (y-axis, ↓ lower is better). ExDBSCAN achieves perfect validity with superior proximity. ExDBSCAN Random and GS-Direct show worse proximity. Surrogate-based methods and BayCon show poor validity (< 50%) due to lack of alignment with DBSCAN model.



**Figure 3: Diversity.** Cumulative percentage of successful queries (x-axis) vs. diversity (y-axis, higher is better). ExDBSCAN achieves highest diversity while maintaining perfect validity, balancing proximity and diversity in its physics-inspired approach.

continuous probability functions. Its optimisation procedure struggles to navigate DBSCAN’s discrete assignment space, meaning that it fails to find any valid transition to counterfactuals.

**Proximity Performance.** Among methods achieving high validity, ExDBSCAN demonstrates superior proximity. ExDBSCAN’s curve lies consistently below both ExDBSCAN Random and DiCE surrogate, indicating that ExDBSCAN produces counterfactuals closer to the original instances. This advantage arises directly from ExDBSCAN’s physics-inspired optimisation approach where the spring-like attraction term in Eq. 4 explicitly biases the selection

toward core points with a minimal weighted-path distance to the explained point  $p$ . By construction, ExDBSCAN’s algorithm initialises with the nearest core points (Prop. 1) and incrementally selects additional cores that minimise total system energy, balancing proximity against diversity constraints.

*ExDBSCAN Random*, while achieving perfect validity, notably shows worse proximity because it samples core points uniformly at random from the target cluster without considering their distance to point  $p$ . This demonstrates that the structured core-point selection in ExDBSCAN is in fact essential for proximity optimisation. *DiCE surrogate* achieves reasonable proximity. However, its

optimisation approach does not take into consideration DBSCAN’s density-connectivity structure, and thereby lacks a mechanism to select the most proximal valid points once multiple candidates are found. Other *surrogate-based* methods like *BayCon*, when they do succeed, often produce counterfactuals with poor proximity values. This occurs because these methods must explore broadly to find any valid points at all, given their poor alignment with DBSCAN’s true decision boundaries. Their successful counterfactuals tend to be those that happened to lie near large, easily-discoverable dense regions rather than the minimal changes that *ExDBSCAN* identifies through its density-aware optimisation. The per-dataset tables showing raw proximity values are presented in the extended version in Table 2 [21].

## 4.2 Diversity Evaluation

Fig. 3 shows the corresponding results for the evaluation metric *diversity*. As before, the  $x$ -axis represents the cumulative percentage of queries with successful counterfactual generation, while the  $y$ -axis reports the diversity score of the returned counterfactual set. Higher diversity values indicate more diverse counterfactual solutions, which is the desired outcome. Consequently, methods whose curves lie higher achieve greater diversity among the generated counterfactuals at the same level of validity, while curves that extend further to the right indicate successful counterfactual generation for a larger fraction of queries.

**Diversity Performance.** *ExDBSCAN* achieves the highest diversity among all methods, with its curve lying consistently above all baselines. This superior diversity is because of its electrostatic-like repulsion mechanism (Eq. 2) which explicitly enforces separation between selected core points. The repulsion force scales with the inverse of the weighted shortest-path distance  $D(V_i, V_j)$  in the cluster’s density-connectivity graph  $G$ , ensuring that selected counterfactuals are spread throughout the cluster structure rather than clustered together. Critically, this diversity measure respects DBSCAN’s notion of similarity. If points appear close in Euclidean space, they may be far apart in terms of density-connectivity, and *ExDBSCAN*’s graph-based distance ensures that selected counterfactuals represent actually distinct alternatives. We illustrate this further in the extended version [21]) *ExDBSCAN Random*, despite having perfect validity, does not match *ExDBSCAN*’s performance on diversity. Its random sampling from core points produce spread-out selections when the cluster contains many spread-out cores. However, this comes at the cost of significantly worse proximity, as seen in Fig. 2, highlighting the proximity-diversity trade-off that *ExDBSCAN* explicitly optimises for. The competitive diversity of this baseline confirms that repulsion-based core selection is an effective diversity strategy; *ExDBSCAN* goes further by systematically achieving this while simultaneously optimising proximity.

*GS-Direct* shows moderate diversity, considerably lower than both *ExDBSCAN* and *ExDBSCAN Random*. *GS-Direct*’s expanding sphere strategy tends to find counterfactuals along similar search directions, particularly when the original point is near a cluster boundary. Without an explicit diversity mechanism, *GS-Direct*’s multiple counterfactuals often represent minor variations along the same transition path instead of truly alternative routes into the target cluster.

The poor diversity performance of the surrogate-based methods and *BayCon*, when they at all find valid counterfactuals, reflects that their fundamental approach is ill-aligned with DBSCAN. These methods optimise against surrogate decision boundaries that do not capture DBSCAN’s density-connectivity structure. Even when they incorporate diversity mechanisms (e.g., *DiCE*’s determinantal point process), the diversity is measured in Euclidean space or with respect to the surrogate’s feature space, but not DBSCAN’s density-connected distance that is the model to be explained here. Consequently, their counterfactuals may appear diverse under their internal metrics while being related or simply invalid from DBSCAN’s perspective. The full per-dataset results are presented in the extended version in Table 3.

**Balancing Proximity and Diversity.** The results from Fig. 2 and 3 demonstrate *ExDBSCAN*’s success in optimising the inherent trade-off between proximity and diversity. By modelling this balance as a physical energy minimisation problem (Eq. 3), *ExDBSCAN* achieves superior performance on both metrics simultaneously. *ExDBSCAN* generates counterfactuals that are both close to the original instance (enabling realistic, minimal changes) and diverse from one another (providing multiple distinct alternatives). These empirical results validate that our physics-inspired approach for *ExDBSCAN* is effective for achieving multi-objective optimisation in this counterfactual generation problem. Thus, our counterfactuals truly support “what-if” scenarios even in the absence of differentiable loss functions or class memberships required for existing methods.

## 5 Conclusion and Future Work

We propose *ExDBSCAN*, a novel counterfactual method specifically designed for density-based clustering. *ExDBSCAN* addresses the interpretability gap in DBSCAN by generating noise-to-cluster and cluster-to-cluster counterfactual explanations with theoretical validity guarantee. Our novel physics-inspired optimisation approach models the proximity-diversity trade-off as an energy minimisation problem, taking into account the structure of the cluster.

Given experiments on 30 tabular datasets, *ExDBSCAN* demonstrates superior performance across all evaluation metrics, generating more proximal counterfactuals than all baselines, and producing more diverse counterfactual sets, while achieving perfect validity.

In future work, *ExDBSCAN* could be extended to incremental or streaming scenarios with evolving clusters; particularly relevant to applications in continuous monitoring systems, fraud detection or real-time anomaly detection, where data arrive incrementally and cluster assignments may change over time.

Investigation into more sophisticated optimisation strategies beyond our greedy solution to the NP-hard energy minimisation problem might further improve counterfactual quality.

## Acknowledgments

This work was partially funded by the Pioneer Centre for AI, DNRF grant number P1, partially supported by project W2/W3-108 Initiative and Networking Fund of the Helmholtz Association and part-funded by the Marie Skłodowska-Curie Doctoral Network RELAX-DN, funded by EU under Horizon Europe 2021-2027 FP Grant Agreement nr. 101072456 ([www.relax-dn.eu/](http://www.relax-dn.eu/)).

## References

- [1] Homayun Afrabandpey and Michael Spranger. 2022. Feasible and Desirable Counterfactual Generation by Preserving Human Defined Constraints. arXiv:2210.05993 [cs.LG] <https://doi.org/10.48550/arXiv.2210.05993>
- [2] Tommaso Amico, Pernille Matthews, Lena Krieger, Arthur Zimek, and Ira Assent. 2026. DCFO: Density-Based Counterfactuals for Outliers. In *KDD (1)*. 37–48.
- [3] Fabrizio Angiulli, Fabio Fassetto, Simona Nisticò, and Luigi Palopoli. 2023. Counterfactuals Explanations for Outliers via Subspaces Density Contrastive Loss. In *DS (Lecture Notes in Computer Science)*. Springer, Porto, Portugal, 159–173.
- [4] Sayan Bandyapadhyay, Fedor V. Fomin, Petr A. Golovach, William Lochet, Nidhi Purohit, and Kirill Simonov. 2023. How to find a good explanation for clustering? *Artif. Intell.* 322 (2023), 103948.
- [5] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *KDD*. ACM, San Francisco, CA, USA, 785–794.
- [6] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *KDD*. AAAI Press, Portland, Oregon, USA, 226–231.
- [7] Ravi B. D. Figueiredo and Hugo Abreu Mendes. 2024. Analyzing Information Leakage on Video Object Detection Datasets by Splitting Images Into Clusters With High Spatiotemporal Correlation. *IEEE Access* 12 (2024), 47646–47655.
- [8] Ziad Francis, Carmen Villagrasa, and Isabelle Clairand. 2011. Simulation of DNA damage clustering after proton irradiation using an adapted DBSCAN algorithm. *Comput. Methods Programs Biomed.* 101, 3 (2011), 265–270.
- [9] Riccardo Guidotti. 2024. Counterfactual explanations and how to find them: literature review and benchmarking. *Data Min. Knowl. Discov.* 38, 5 (2024).
- [10] David Halliday, Robert Resnick, and Jearl Walker. 2013. *Fundamentals of physics*. John Wiley & Sons, Hoboken, NJ, USA.
- [11] Kentaro Kanamori, Takuya Takagi, Ken Kobayashi, and Hiroki Arimura. 2020. DACE: Distribution-Aware Counterfactual Explanation by Mixed-Integer Linear Optimization. In *IJCAI*. ijcai.org, Virtual, 2855–2862.
- [12] Amir-Hossein Karimi, Bernhard Schölkopf, and Isabel Valera. 2021. Algorithmic Recourse: from Counterfactual Explanations to Interventions. In *FACt*. ACM, Toronto, Canada, 353–362.
- [13] Antonia Karra, Georgios Vardakas, Evaggelia Pitoura, and Aristidis Likas. 2025. Generating Counterfactual Explanations for Clustering Models Based on Their Equivalence to Classification Models. In *AIAl (1) (IFIP Advances in Information and Communication Technology)*. Springer, Crete, Greece, 85–100.
- [14] Been Kim, Oluwasanmi Koyejo, and Rajiv Khanna. 2016. Examples are not enough, learn to criticize! Criticism for Interpretability. In *NIPS*. Curran Associates, Inc., Red Hook, NY, USA, 2280–2288.
- [15] Alex Kulesza and Ben Taskar. 2012. Determinantal Point Processes for Machine Learning. *Found. Trends Mach. Learn.* 5, 2-3 (2012), 123–286.
- [16] Bum Chul Kwon, Ben Eysenbach, Janu Verma, Kenney Ng, Christopher deFilippi, Walter F. Stewart, and Adam Perer. 2018. Clustervision: Visual Supervision of Unsupervised Clustering. *IEEE Trans. Vis. Comput. Graph.* 24, 1 (2018), 142–151.
- [17] Thibault Laugel, Adulam Jeyasothy, Marie-Jeanne Lesot, Christophe Marsala, and Marcin Detyniecki. 2023. Achieving Diversity in Counterfactual Explanations: a Review and Discussion. In *FACt*. ACM, Chicago, IL, USA, 1859–1869.
- [18] Thibault Laugel, Marie-Jeanne Lesot, Christophe Marsala, Xavier Renard, and Marcin Detyniecki. 2019. The Dangers of Post-hoc Interpretability: Unjustified Counterfactual Explanations. In *IJCAI*. ijcai.org, Macao, China, 2801–2807.
- [19] Dan Ley, Umang Bhatt, and Adrian Weller. 2022. Diverse, Global and Amortised Counterfactual Explanations for Uncertainty Estimates. In *AAAI*. AAAI Press, Vancouver, BC, Canada, 7390–7398.
- [20] Bingqiao Luo, Zhen Zhang, Qian Wang, Anli Ke, Shengliang Lu, and Bingsheng He. 2025. AI-powered Fraud Detection in Decentralized Finance: A Project Life Cycle Perspective. *ACM Comput. Surv.* 57, 4 (2025), 96:1–96:38.
- [21] Pernille Matthews, Lena Krieger, Tommaso Amico, Arthur Zimek, Thomas Seidl, and Ira Assent. 2026. ExDBSCAN: Explaining DBSCAN with Counterfactual Reasoning - Additional Material. <https://arxiv.org/abs/2605.30225>
- [22] Nijat Mehdiyev, Maxim Majlatow, and Peter Fettke. 2024. Counterfactual Explanations in the Big Picture: An Approach for Process Prediction-Driven Job-Shop Scheduling Optimization. *Cogn. Comput.* 16, 5 (2024), 2674–2700.
- [23] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artif. Intell.* 267 (2019), 1–38.
- [24] Christoph Molnar. 2020. *Interpretable machine learning*. Lulu.com, Morrisville, NC, USA. <https://christophm.github.io/interpretable-ml-book/>
- [25] Michal Moshkovitz, Sanjoy Dasgupta, Cyrus Rashtchian, and Nave Frost. 2020. Explainable k-Means and k-Medians Clustering. In *ICML (Proceedings of Machine Learning Research)*. PMLR, Virtual, 7055–7065.
- [26] Ramaravind Kommiya Mothilal, Amit Sharma, and Chenhao Tan. 2020. Explaining machine learning classifiers through diverse counterfactual explanations. In *FAT\**. ACM, Barcelona, Spain, 607–617.
- [27] Davoud Moulavi, Pablo A. Jaskowiak, Ricardo J. G. B. Campeglio, Arthur Zimek, and Jörg Sander. 2014. Density-Based Clustering Validation. In *SDM*. SIAM, Philadelphia, PA, USA, 839–847.
- [28] Francisco Parreño, Ramón Alvarez-Valdés, and Rafael Martí. 2021. Measuring diversity. A review and an empirical analysis. *Eur. J. Oper. Res.* 289, 2 (2021), 515–532.
- [29] Rafael Poyiadzi, Kacper Sokol, Raúl Santos-Rodríguez, Tijl De Bie, and Peter A. Flach. 2020. FACE: Feasible and Actionable Counterfactual Explanations. In *AIES*. ACM, New York, NY, USA, 344–350.
- [30] Piotr Romashov, Martin Gjoreski, Kacper Sokol, Maria Vanina Martinez, and Marc Langheinrich. 2022. BayCon: Model-agnostic Bayesian Counterfactual Generator. In *IJCAI*. ijcai.org, Vienna, Austria, 740–746.
- [31] Christian A. Scholbeck, Henri Funk, and Giuseppe Casalicchio. 2023. Algorithm-Agnostic Feature Attributions for Clustering. In *xAI (1)*. Springer, 217–240.
- [32] Aurora Spagnol, Kacper Sokol, Pietro Barbiero, Marc Langheinrich, and Martin Gjoreski. 2024. Counterfactual Explanations for Clustering Models. arXiv:2409.12632 [cs.LG] <https://arxiv.org/abs/2409.12632>
- [33] Ilija Stepin, José Maria Alonso, Alejandro Catalá, and Martin Pereira-Fariña. 2021. A Survey of Contrastive and Counterfactual Explanation Generation Methods for Explainable Artificial Intelligence. *IEEE Access* 9 (2021), 11974–12001.
- [34] Joaquin Vanschoren, Jan N. van Rijn, Bernd Bischl, and Luis Torgo. 2013. OpenML: networked science in machine learning. *SIGKDD Explor.* 15, 2 (2013), 49–60.
- [35] Georgios Vardakas, Antonia Karra, Evaggelia Pitoura, and Aristidis Likas. 2025. Counterfactual Explanations for k-Means and Gaussian Clustering. In *ICTAI*. IEEE, Athens, Greece, 977–983.
- [36] Sandra Wachter, Brent D. Mittelstadt, and Chris Russell. 2017. Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR. arXiv:1711.00399 [cs.AI] <https://arxiv.org/abs/1711.00399>
- [37] Shoki Yamao, Ken Kobayashi, Kentaro Kanamori, Takuya Takagi, Yuichi Ike, and Kazuhide Nakata. 2024. Distribution-Aligned Sequential Counterfactual Explanation with Local Outlier Factor. In *PRICAI (1) (Lecture Notes in Computer Science)*. Springer, Kyoto, Japan, 243–256.
- [38] Wenzhuo Yang, Jia Li, Caiming Xiong, and Steven C. H. Hoi. 2022. MACE: An Efficient Model-Agnostic Framework for Counterfactual Explanation. arXiv:2205.15540 [cs.LG] <https://arxiv.org/abs/2205.15540>
- [39] Yining Yuan, Kevin McAreevey, Shujun Li, and Weiru Liu. 2024. Multi-Granular Evaluation of Diverse Counterfactual Explanations. In *Proceedings of the 16th International Conference on Agents and Artificial Intelligence, ICAART 2024, Volume 2, Rome, Italy, February 24-26, 2024*. Ana Paula Rocha, Luc Steels, and H. Jaap van den Herik (Eds.). SCITEPRESS, Rome, Italy, 186–197. doi:10.5220/0012349900003636
- [40] Peng Zhou, Qihui Tong, Shiji Chen, Yunyun Zhang, and Xindong Wu. 2025. EACE: Explain Anomaly via Counterfactual Explanations. *Pattern Recognit.* (2025).

## A Appendix

### A.1 LLM statement

No generative AI tools were used in the writing of the paper and at any stage of this research.

### A.2 Algorithm Pseudocode

To complement Section 3.3, we include full pseudocode for the two ExDBSCAN components that balance proximity and diversity:

- (1) Selection of  $k$  reference core points via greedy minimisation of energy function in Eq. 3.
- (2) Generation of counterfactual examples from the selected core points using Eq. 4.

**A.2.1 Greedy Selection of Core Points.** Given a point  $p$  to be explained and a target cluster, the first step of ExDBSCAN is to select a set of  $k$  core points whose  $\epsilon$ -neighbourhoods will serve as the basis for constructing counterfactuals.

The energy of a candidate set  $C'$  is:

$$E_{C'} = \sum_{V_i \in C'} \sum_{V_j \in C': j > i} \frac{1}{\mathcal{D}(V_i, V_j)} + \sum_{V_i \in C'} d^2(p, V_i).$$

The optimisation problem in Eq. 3 is NP-hard (Proposition 2), so we use a greedy approximation. For  $k = 1$ , this procedure returns the optimal solution.

**A.2.2 Constructing Counterfactuals.** Once the set of reference core points  $C' = \{q_1, \dots, q_k\}$  has been selected, each counterfactual  $p'$  is

**Algorithm 1** Greedy Selection of Core Points

**Require:** Point to explain  $p$ ; set of core points  $V$ ; number of counterfactuals  $k$ ; weighted graph distance  $\mathcal{D}(\cdot, \cdot)$ ; feature-space distance  $d(\cdot, \cdot)$ ; target cluster  $t$ .

**Ensure:** Set  $C'$  of  $k$  selected core points

```

1:  $C' \leftarrow \emptyset$ 
2:  $V^t = \{V_j : V_j \in t\}$ 
3: while  $|C'| < k$  do
4:    $\text{best\_core} \leftarrow \text{None}$ ,  $\text{best\_energy} \leftarrow +\infty$ 
5:   for each core point  $V_j \in V \setminus C'$  do
6:      $S \leftarrow C' \cup \{V_j\}$ 
7:      $E_S \leftarrow 0$   $\triangleright$  Repulsion term (graph distances)
8:     for all pairs  $(V_i, V_\ell)$  in  $S$  with  $\ell > i$  do
9:        $E_S \leftarrow E_S + \frac{1}{\mathcal{D}(V_i, V_\ell)}$ 
10:    end for  $\triangleright$  Attraction term (feature-space distance)
11:    for all  $V_i \in S$  do
12:       $E_S \leftarrow E_S + d^2(p, V_i)$ 
13:    end for
14:    if  $E_S < \text{best\_energy}$  then
15:       $\text{best\_energy} \leftarrow E_S$ 
16:       $\text{best\_core} \leftarrow V_j$ 
17:    end if
18:  end for
19:   $C' \leftarrow C' \cup \{\text{best\_core}\}$ 
20: end while
21: return  $C'$ 

```

**Algorithm 2** Generate Counterfactuals from Selected Core Points

**Require:** Point  $p$ ; selected core points  $C' = \{q_1, \dots, q_k\}$ ; DBSCAN radius  $\varepsilon$ ; feature-space metric  $d(\cdot, \cdot)$

**Ensure:** Set  $C$  of  $k$  counterfactuals

```

1:  $C \leftarrow \emptyset$ 
2: for all  $q_j \in C'$  do
3:    $d_{pq} \leftarrow d(p, q_j)$   $\triangleright$  By construction,  $d_{pq} > 0$  since  $p$  and  $q_j$ 
   belong to different clusters
4:    $\text{direction} \leftarrow (q_j - p)/d_{pq}$ 
5:    $p' \leftarrow p + (d_{pq} - \varepsilon) \cdot \text{direction}$   $\triangleright$  Eq. 4
6:    $C \leftarrow C \cup \{p'\}$ 
7: end for
8: return  $C$ 

```

placed inside the  $\varepsilon$ -neighbourhood of its corresponding core point  $q$ . To maximise proximity to the original point  $p$ , we move away from  $p$  towards  $q$  and stop at distance  $\varepsilon$  from  $q$ . This corresponds exactly to Eq. 4 in the main text. Because each constructed point lies within  $\varepsilon$  of a core point in the target cluster, Theorem 3.1 guarantees that all counterfactuals are valid. Algorithm 2 summarises the procedure through pseudocode. Because each  $p'$  lies within distance  $\varepsilon$  of a core point in the target cluster, the construction ensures all counterfactuals are valid.

**A.2.3 Target Cluster Not Specified.** Pseudocode 1 requires the target cluster in input. The user could prefer not to specify a certain target cluster, e.g. if the user wants to convert a noise point to a generic non-noise instance. In this case, pseudocode 1 would need

to be slightly changed on line 8. Specifically, instead of adding a repulsion term for each pair of core points already selected unconditionally, we add it only if they belong to the same cluster. Points belonging to different cluster are indeed already diverse by construction and they are not connected through any density connected path according to DBSCAN's algorithm.

**A.2.4 Complexity Analysis.** Given a DBSCAN clustering as input (with core points and weighted graph distances), ExDBSCAN has the following computational complexity. Core point selection requires  $O(k^3 * |V^t|)$ , construction of CEs requires  $O(kn)$ , and pre-computing distances requires  $O(n|V^t| + |V^t||E|)$ , where  $n$  denotes the number of features,  $V^t$  the set of core points in target cluster  $t$ ,  $E$  the edges in connectivity graph, and  $k$  the number of CEs.

Validity is guaranteed by Theorem 3.1 and requires no additional verification step. Regarding scalability: high-dimensionality increases  $n$ , while large-scale datasets and the neighbourhood radius  $\varepsilon$  do not directly influence scalability, but implicitly in  $|V^t| \gg N$  and  $E$ , both of which scale linearly. Only  $k$  enters cubically, however it is a small constant in practice, e.g., [26] uses 3-5. Overall, runtime remains tractable in high-dimensional and large scale settings.

**A.3 NP-hardness**

Proposition 2, states that the optimisation problem in Equation 3 is *NP-hard*, we present here a more extended proof.

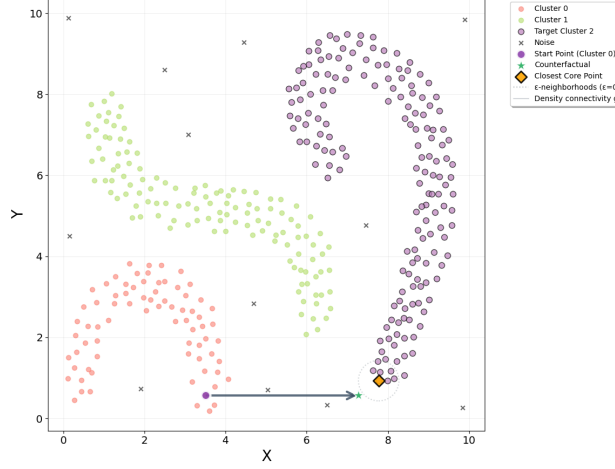
**PROOF.** Consider the energy landscape in Equation 2, the electrostatic term reads (in contracted notation)  $\sum_{i,j>i} \frac{1}{\mathcal{D}(V_i, V_j)}$ , which needs to be minimised to find the optimal energy configuration. Minimising the sum of the energy terms is equivalent to maximise their negative. Thus, the problem can be mapped to a maximum diversity problem where the object summed is the inverse distance, MDP optimisation is known to be NP-hard [28]. Consider now adding the spring terms, considering Equation 2 in full. We can again map the minimisation problem to a maximisation one by taking the negative energy. If this new optimisation problem would be solvable in polynomial time, we could move the spring constants (which we consider equal to 1) to 0. In this limit, the MDP could be also solved in polynomial time, contradicting its NP-hardness. ■

**A.4 Approximation Quality of CE Selection**

The energy minimization problem (Eq. (3)) is NP-hard and solved via a greedy strategy. To assess approximation quality, we additionally evaluate a local search variant (best-improvement swap initialized from greedy solution) and an exact solver (Gurobi MIQP on Eq. (3)), across 10 datasets, 100 query points and  $k \in \{4, 10\}$ .

For  $k = 4$ , the greedy solution matches the exact optimum in 87% cases (mean ratio 1.0014, worst-case 1.05, 98% within 1%). Local search initialized from the greedy solution converges to the exact optimum in 100% cases, requiring a mean of only 0.2 swaps, which confirms that greedy is near-locally-optimal. For  $k=10$ , the gap shrinks, greedy achieves 90% exact matches (mean ratio 1.0005, worst-case 1.02) and local search reaches 99% with 0.1 mean swaps. Across both values of  $k$ , all cases fall within 5% of optimum.

To assess whether objective differences translate to differences in the selected sets, we report mean Jaccard overlap between the greedy and exact solutions, for  $k = 4$  0.94 and 0.97 for  $k = 10$ . Even



**Figure 4: Illustration of non-actionability.** Coloured points denote cluster assignments. The x-axis shows the actionable feature and the y-axis the non-actionable feature. The orange rhombus marks the closest core point within  $\varepsilon$  in the non-actionable direction from the point to explain (purple), and the green star indicates the counterfactual.

in cases where objective values differ, the selected CE sets are nearly identical. We attribute this to the sensitivity of the repulsion term to close pairs: once greedy avoids selecting nearby points, further optimization yields only marginal improvements.

## A.5 Non-Actionable Features

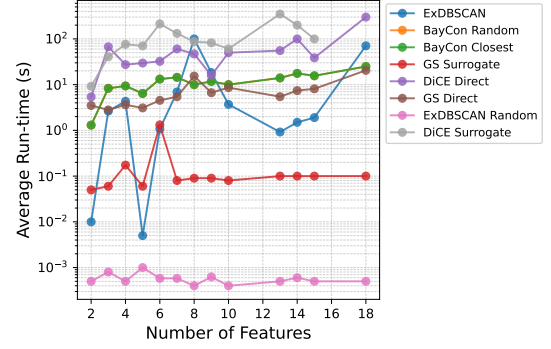
Actionability ensures that the generated CEs are actionable, i.e., a user can actually realistically change the features. Non-actionable features such as genetic information or inherent personal attributes, make explanations unrealistic and prevent human stake-holders to simulate and apply the proposed changes.

**A.5.1 Non-Actionable Features in ExDBSCAN.** To account for non-actionable features, we restrict the set of core points to consider, when identifying the set  $C'$ . The constrained set of core points includes all core points that are not further than  $\varepsilon$  away from the original point when measuring distances in the subspace defined by the non-actionable features  $\mathbb{R}^{n'}$ , where  $n' < n$  denotes the number of non-actionable features. In this way, we take into consideration only core points whose  $\varepsilon$  neighbourhood is reachable without changing any of the non-actionable features.

The undirected weighted graph  $G$  is then built and the set of reference core points chosen  $C'$  as described in Section 3.3.

When moving from the set of core points  $C'$  to the set of counterfactuals  $C$ , it is not guaranteed that we can move in the direction of the point to explain  $p$ , as we now have the constraint given by the non-actionable features. Therefore, we move toward point  $p$  only in the subspace  $\mathbb{R}^{n-n'}$  formed by the actionable features, still putting the counterfactual  $\varepsilon$  away from the reference core point, which is always possible given our core point filtering procedure.

**A.5.2 Real-world Constraints.** Several strategies can account for real-world constraints. First, if the non-actionable features include



**Figure 5: Lineplot showing the runtime in seconds (y-axis, log-scaled) and number of features (x-axis).**

categorical features, we suggest changing the metric used in DBSCAN and ExDBSCAN; for details on metric choice, see Section 3.3.

Other constraints, i.e., monotonicity or feature bounds can be easily incorporated through further filtering of the core point set. Instead of restricting the set of reference core points to actionable dimensions only, we can additionally restrict this to be a specific direction or range depending on the constraints defined by the user, which allows incorporating known domain knowledge.

**A.5.3 Non-Actionable features.** We here evaluate ExDBSCAN in the setting where a subset of features are non-actionable. We utilise the experimental pipeline used in Section 4. How many features are non actionable is a parameter chosen at random between one 1 and  $n/2$  features, with  $n$  the total number of features.

**A.5.4 Non-Actionable Results.** We present an experimental analysis benchmarking ExDBSCAN when a subset of features is non-actionable, i.e. cannot be modified by the counterfactual method. We follow the experimental setting of Section 4, but additionally randomly select a subset of features, with cardinality between 0 and half of the features, as non-actionable.

Empirical percentile curves, visualising the distribution of proximity while simultaneously reflecting method failure rates are presented in Figure 6 ([21]); diversity information is instead visualised in Figure 7 ([21]). Results mirror the ones presented in Section 4, with ExDBSCAN outperforming competitors in validity, proximity and diversity. Full per-dataset tables are presented in the extended version in Tables 7, 6, and 5.

## A.6 Run-Time Analysis

Figure 5 shows average runtime as a function of the number of features. ExDBSCAN scales linearly with dimensionality and remains competitive throughout. ExDBSCAN Random is consistently fastest given its simple sampling heuristic, while ExDBSCAN and Growing Spheres are the next quickest. DiCE variants and BayCon are substantially slower across all settings, with DiCE Surrogate and DiCE Direct ranking as the heaviest methods computationally. Cumulative runtime curves and full per-dataset runtime tables are provided in the extended version [21]