Advanced Engineering Informatics 48 (2021) 101261

Contents lists available at ScienceDirect

Advanced Engineering Informatics

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

Common design structures and substitutable feature discovery in CAD databases

Gokula Vasantha [a,*], David Purves [b], John Quigley [b], Jonathan Corney [c], Andrew Sherlock [c], Geevin Randika [c]

[a] School of Engineering and the Built Environment, Edinburgh Napier University, EH10 5DT, Scotland, UK
[b] Department of Management Science, University of Strathclyde, Glasgow G1 1XQ, Scotland, UK
[c] School of Engineering, University of Edinburgh, EH8 9YL, UK

ARTICLE INFO

Keywords:
3D feature recognition
CAD feature reuse
Common design structure
Data mining
Design search
Substitutable feature

ABSTRACT

It has been widely reported that the reuse of previously created components, or features, in new engineering designs will improve the efficiency of a company's product development process. Although the reuse of engineering components has established metrics and methodologies, the reuse of specific design features (e.g. stiffening ribs, hole patterns or lubrication grooves, etc.) has received less attention in the literature. Typically, researchers have reported approaches to partial design reuse that identify patterns predominately in terms of geometrically similar shapes (i.e. a set of features) whose elements are adjacent, cohesive, and decoupled from the overall form of a component.

In contrast, this paper defines a common design structure (CDS) as collections of frequently occurring features (e.g. holes) with common parametric values (e.g. diameters) in a CAD database (irrespective of their locations or spatial connectivity between other features on a component). By exploiting the established data-mining technology of association rules and item-sets the authors show how CDSs can be efficiently computed for hundreds of 3D CAD models. A case study, with hole data extracted from a publicly available dataset of hydraulic valves, is presented to illustrate how item-sets associated with CDS can be computed and used to support predictive design by identifying potentially 'substitutable features' during an interactive design process. This is done using a combination of association rules and geometric compatibility checks to ensure the system's suggestion are implementable. The use of the Kullback–Leibler divergence to assess the degree of similarity between components is identified as a crucial step in the process of identifying the "best" suggestions. The results illustrate how the prototype implementation successfully mines the CDSs and identifies substitutable hole features in a dataset of industrial valve designs.

1. Introduction

Many large corporations have product portfolios that span multiple generations and physical sites. However, the tools for analysing the contents of this rich and varied dataset are relatively limited. Studies have suggested that a new product's development time can be reduced by up to 80% by effective design reuse [14]. But although the reuse of engineering components has established metrics and methodologies, the reuse of specific design features (e.g. stiffening ribs, hole patterns or lubrication grooves, etc.) has received less attention in the literature. The nature of such design features varies from product to product but can be defined generically as a "portion of the product geometry that is of design significance". Zhang et al. [25] mentioned that design features commonalities are usually identified manually using the tacit knowledge of experienced engineers; a process which lacks the speed and detail required for digital management of design portfolios. This paper describes a new approach to reusing common shape features from 3D CAD models. Instead of searching CAD models to identify instances of particular patterns of local geometry (i.e. geometric feature recognition), the aim is to find shared patterns in the feature content of multiple models. In other words, we are seeking to recognise sets of features that frequently occur together in a CAD database.

* Corresponding author.
E-mail addresses: G.Vasantha@napier.ac.uk (G. Vasantha), David.Purves@strath.ac.uk (D. Purves), J.Quigley@strath.ac.uk (J. Quigley), J.R.Corney@ed.ac.uk (J. Corney), A.Sherlock@ed.ac.uk (A. Sherlock).

https://doi.org/10.1016/j.aei.2021.101261
Received 12 September 2020; Received in revised form 29 December 2020; Accepted 8 February 2021
Available online 18 March 2021

There are various terminologies used in the literature to refer to a common shape in 3D CAD models, such as: common design structures [17], partial shapes [3,4,22], design pattern [4], interacting features [21] and subpart [13,12]. In this paper, we have adopted the term 'Common Design Structure (CDS)' to mean a set of shape features frequently observed in a given collection of 3D CAD models. In the literature CDSs have been used for automating design knowledge discovery and design feature library customisation and standardisation [17]. Feature-based CDS discovery are also important because of their central role in many commercial user interface paradigms for engineering CADs (e.g. pitch circle patterns of holes) and their ability to express design semantics.

So, a CDS is composed of a set of features that frequently occur in a CAD database. More formally a CDS is defined as a problem of frequent substructure discovery that appears above a given frequency threshold value in a set of 3D models [17]. In addition to frequency, there are many parameters used to define meaningful CDS such as adjacent, cohesive and decoupled features. However, adding more parameters restricts the number of CDSs extracted and consequently the number of potential CDS identified.

The approach presented here expands the notion of what constitutes a reusable CAD feature to include frequently occurring patterns that might not be spatially adjacent. So, while the literature has focused mainly on defining reusable regions using concepts of cohesion, coupling, complexity, information richness, frequency and interaction between features, this paper defines reusable design patterns in terms of occurrence rather than location. Consequently, the following statements were formulated to characterise this new approach: (i) all CAD features have the potential for reuse irrespective of their connectedness, dependency and complexity; (ii) commonly occurring sets of CAD features in the database have more potential to be reused together; (iii) similarity in parametric values that define features play a vital role in extracting reusable design structure and (iv) a reusable design structure can have a set of features that could satisfy multiple functions in a product. In these postulates, feature size will be provided higher importance because a change to feature size could necessitate alterations to the manufacturing process.

The presented approach differs significantly from previously reported approaches to both component reuse (typically enabled by a similarity assessment of a part's global shape) and feature reuse (commonly implemented as a search for a predefined pattern of faces and edges in a 3D model's topology and geometry). Rather than search for matches to a target shape, the authors compute lists (i.e. sets) of commonly occurring entities or features in the models of a CAD database. This approach means that what constitutes a "frequently used feature" does not have to be predefined but instead can be "discovered" by analysis of the lists. To do this we used the concept of item-sets, first reported in [1] for the identification of frequent patterns in the content of supermarket shopping trolleys. By defining the problem in terms of the items and item-sets the software utilities developed to support "Association Rule Mining" by computer science researchers, can be exploited to minimise both implementation and run times. Association Rule Mining (ARM) seeks relationships between common sets of items (itemsets) in databases. The relationships can be positive (e.g. if someone buys bread, they will probably buy butter) or negative (e.g. if a customer buys vegetarian burgers, they will not buy sausages). When two itemsets differ by only one item, that item is said to be substitutable (e.g. cake and croissant are substitutable in a frequent pair of items with coffee).

This research work aims to assess the potential of Association Rule Mining to support: CDS extraction in CAD models and find 'substitutable features' within emerging CDSs. The potential of extracting substitutable CAD features from CDSs is analysed for the first time in this paper. Informally 'substitutable features' can be described as being analogous to synonyms for words in written languages. Similarly, 'substitutable features' can potentially be exchanged for other features that are judged to be equivalent because of the frequency with which they occur within the set of features a design contains. This provides a method of data mining design options for designers to choose, or ignore, as appropriate (like predictive text systems on mobile phones). However, as with predictive text, it is possible for the system to produce suggestions that are semantically inappropriate.

So, to increase the likelihood of substitutable features being functionally appropriate to an engineering design, five conditions were used to filter the candidate substitutable features identified by association rule algorithms. These are:

(i) Substitutable features never co-exist together in common design structures;
(ii) Feature occurrences in the component remain the same between substitutable features;
(iii) Substitutable features are associated with common design structures that differ by only one feature;
(iv) The similarity score between components that share substitutable features is close; and
(v) The defining parametric value of substitutable features is within a close range.

These conditions are used to ensure that the incorporation of a substitutable feature in a design does not fundamentally change the shape of a design. This has been recognised as an important issue in previous studies, for example [14] indicated that 48% of the surveyed researchers highlight the major challenge for design reuse is inflexible design models that fail after changes.

A CAD dataset with nearly 2000 models of hydraulic valve assemblies was downloaded from an online catalogue to investigate the proposed approach. In this paper, the focus is to understand CDSs identified with the hole features found on each component in this dataset. Considering only hole features permitted an in-depth analysis on a single feature type rather than focusing on the breadth of feature types (i.e. slots, pockets, keyways etc) that could have been examined. The results obtained using the extracted hole features illustrate the successful discovery of CDSs and substitutable hole features in an industrial valve design dataset. The following sections of the paper discuss: previously reported work in CDS discovery; detail the approach used for extracting and validating item-sets associated with CDS patterns; describe the valve design dataset and the CDSs identified by 'Association Rule Mining', elaborate the identification of substitutable hole features, and concludes with a discussion and future work.

2. Related literature

This section focuses on the literature related to the discovery of design structures in 3D CAD B-rep models. The goal of the survey was to establish how the previous researchers had approached three important questions: 1) How to define a design structure in a CAD model? 2) What mechanisms are used to extract reusable design structure? and 3) How the design structure is assessed and applied? The following subsections summarise the answers to these questions found in the reported work.

2.1. Common design structure definition and associated properties

Commonly, Common Design Structure Discovery (CDSD) is defined to identify local (i.e. regional) structures shared by multiple models [17,3,4,25,26,23]. Ma et al. [17] defined the common design structure discovery problem as follows: "Given a set of 3D models M = {m_1, m_2, m_3, …} and a partial model m' appears f_i times as a substructure of m_i in the dataset M and $\sum_{i=1}^{|M|} f_i \geq \zeta; f_i = \{0,1\}$for a given threshold value ζ, then m' is called a common design structure of M". Although the frequency is an important variable in defining CDS, many other parameters have been defined to identify reusable CDS.

Table 1
Summary and perspective views on parameters to define reusable CSD.

Parameters	CDS Phase	Perspective views of this research
Repetition	CDS extraction	CDS should have repetitive features that are frequently occurring together in the database.
Cohesive		Design features do not have to be cohesive with each other. Features could be distributed across the 3D CAD model to represent a CDS.
Decoupled		Features could be either coupled, or decoupled, since design features are studied individually.
Complexity		All design features have equal importance because of the importance given to reuse merits rather than designing time.
Rich information	Application development	A CDS containing many features should have higher importance relative to simpler ones. However, the clarity to extract design semantics should not be compromised.
Function		A CDS could have collections of features that serve multiple functions instead of a single function.
Substitutable		CDS should support the elicitation of substitutable features generated in the CAD database.
Dependant Intersecting Adjacent		Dependant, intersection and adjacent variables are considered important at the second stage of understanding the characteristics of CDS, but not at the first phase of extracting CDS.
Reusability Compatibility	Application implementation	Ease of reuse and compatible with target design should be factors at the third stage when CDSs are considered for reuse.
Scalability		Adding new features into the CDS is not considered essential. However, understanding the growth of CDS is important.
Maintainability		Possibility of modifying the CDS should not be considered during the CDS extraction process.
Portability		It will be beneficial if the CDS are usable across the different CAD platforms, but it is not essential.
Comprehensibility		CDS should be easy to understand. However, it could have a bit of abstraction to aid creative reuse in CAD modelling.

Bai et al. [3,4] defined a CDS as a set of similar reusable regions in CAD models with the same functions. They used 'cohesive inside (i.e. all the features in the design pattern are connected)', 'decoupled from outside (i.e. independent from the rest of the model)', and 'relative complexity to justify the time spent in design searches and reuse' as reuse criteria to find reusable structures of 3D CAD models. These criteria provided a rational for the proposed CDS having properties of connectedness, independence and association with a single function. The complexity criterion was measured by the number of feature dependency, adjacent or intersecting relationships with other features, and the depth of the feature tree corresponding to the structure.

Bai et al. [4] further expanded the above criteria by arguing that a good design structure should have five specific characteristics (i.e. reusability, scalability, maintainability, comprehensibility and portability) and six conditions (i.e. simple design features, high cohesion, low coupling, moderate complexity, high repetition rate, and rich information). Zhang et al [26] emphasised that a CDS has a lot of versatility and high reuse value since its structure frequently appears in different product models.

Li et al. [16] proposed a geometric reasoning approach to generate hierarchy for B-rep model retrieval. The hierarchical representation retains parent–children-sibling feature relationships that covers geometric and topological information for global and partial retrieval. The representation is multi-resolutional, covering faces to features with partitions and segmentations. They emphasised that reusable structures should be partitioned by faces that share coherent concave or convex adjacency, correlated with feature segmentation, share local topology and shape distribution, and ensure compatibility of feature boundaries between the source feature and the target reuse base model.

Sunil et al. [21] emphasised recognition of interacting features in B-Rep CAD models for identifying design structures. They considered variable topology features and handled adjacent and volumetric feature interactions to provide a single topology interpretation. Table 1 summarises the list of parameters used to define CDS in the literature and defines the context of the work presented in this paper. Among the listed parameters, substitutability can be seen as a novel additional parameter that is not mentioned in the literature. The review also highlights that considering too many parameters is restrictive and hinders the full understanding of potential CDSs in a CAD database. From this unrestricted view, the CDS is defined as a set of frequently occurring parametric design features occurring together in the database that could achieve multiple functions and be located in any parts of a CAD model.

This research work proposes three progressive stages for effective CDS usage: CDS extraction, application development and application implementation. In the CDS extraction process, the emphasis is only given to a set of features that are frequently occurring in the database. Other parameters such as dependency, intersection and adjacent relationships could be chosen based on the type of application in which CDS will be used. In this paper, CDS is used to identify substitutable features in the CAD database. Parameters such as ease of reuse and compatible with target design should be considered at the third stage of application implementation.

2.2. *Approaches to extract common design structure*

The reported steps to extract CDS most commonly involve describing the component with an appropriate representation, clustering similar parts, and extracting CDS based on frequency threshold and other defined attributes. Various graph-based approaches have been reported for both representing components and extracting CDS from B-rep CAD models. Ma et al. [17] defined common design patterns as frequently appearing subgraphs of a model's face adjacency graph (FAG) drawn in a plane. They used FAG to express topologic relations in B-rep model where nodes correspond to faces of the model and edges correspond to adjacency relations between faces of the model. Local shape characteristics such as adjacent faces, geometry type, intersection curve type between two adjacent faces, and the average dihedral angle between the two faces of an edge were attached to its nodes. Each node of a FAG was mapped onto a two-dimensional plane which represent the face's shape characteristics with two coordinates. Frequent subgraphs were discovered by comparing the shape descriptors composed from the point coordinates using Apriori-based graph mining technique. Comparing the shape descriptors composed from the point coordinates avoids using the exact subgraph-isomorphism checking. However, using face information limited the focus on adjacency in extracting CDS. Another limitation is that the descriptive graph code (calculated from the shape parameters) is not guaranteed to distinguish all the shapes in the CAD models.

Bai et al. [3,4] extracted CDS by generating isolated patterns in the extended feature tree from B-rep model, where the nodes of the graph represent design features, the edges of the graph represent the relationship between the design features, and one attribute of each node represents its adjacent and intersecting relationships. The isolated proper subtrees, along with hierarchically described local matching regions, were used for CDS retrieval. The isolated proper subtree was

characterised as having high cohesion and low coupling attributes, and not directly connected to the root of the tree. The extracted subparts were clustered using a graph-oriented, agglomerative, hierarchical clustering algorithm, and final design structures were identified by calculating a complexity score. The complexity score was calculated based on the types of design features and relationships between different design features. One advantage of the approach was that sub-tree, rather than the harder sub-graph, matching could be used to identify frequency information. The consideration of many parameters (such as high cohesion, low coupling and complexity attributes) have limited the understanding of the potential of CDS. The hierarchically described extended feature tree largely relies on adjacency between features, and location points of the features are not considered in the extended feature tree. There is a possibility that a tree may be designed differently from one designer to another for the same model.

Along with CDS extraction in component models, methods have also been reported for extracting CDS from assembly models. Zhang et al. [25] used mating face pairs (MFP) and a generic face adjacency graph (GFAG) to quantitatively describe assembly models and their relationships to extract CDS in assembly models. A MFP is defined as "two faces mating with each other or being contained in a joint, and they must belong to two different parts of an assembly model". The MFP includes low-level face mating constraints such as coincidence, contact, offset, angle, etc. Point classification based on principal curvatures and the dihedral angle was used to capture geometric shape characteristics quantitatively, which describes nodes and arcs in GFAG, respectively. The probability distribution of different types of points was used to describe the faces and the edges. Again the Apriori-based graph mining (AGM) approach was used to extract CDS from those quantitative GFAG assembly representation. Point-based classification to represent geometry is time-consuming and involves some approximation to capture geometric shapes.

Zhang et al. [26] used attribute connection graphs (ACGs) to represent topological information and attributes of parts and connections in assembly models. In an ACG, each node represents a part, and edges represent topological relations. K-means clustering is used to classify the parts and connections based on similarity analysis of different attributes, and fast, frequent subgraph mining used to identify common design structures in assemblies from the ACGs of assemblies transformed into type code graphs. The qualitative type representation could have textual ontological issues such as misinterpretation and difficulty in expanding ontological terminologies.

Wang et al. [23] proposed the generic face adjacency graph to discover the common design structures in assembly models. Shape vectors and link vectors were used to describe quantitatively the part models and mating relationships, respectively. These vectors capture geometrical and topological information of the assembly model. The similar parts and mating relationships were clustered and labelled using distance measure and cosine similarity respectively by a k-means approach. Specifically they used the gSpan algorithm [24] to mine the frequent subgraphs from the clustered graphs that satisfy the given predefined threshold value.

The approaches proposed in the literature are well developed for identifying CDSs based on adjacent faces and features. Attributed face relations from B-rep models, and feature-based graphs of CAD models dominate the representation of adjacent relationships. The relationships between features such as dependent, intersecting and adjacent have played important roles in extracting CDS in the proposed approaches. However, this research does not require that features in a CDS have rigorously defined relationships, instead the authors approach assumes a number of independent features can form CDSs, that are useful for design reuse, if they frequently occur together on parts in the CAD database. Although the tree-based approach is effective in mining CDSs (compared to, say, a subgraph-based approach), the process is still computationally intensive. In contrast the itemset-based approach adopted in this work has, potentially, computationally efficient implementations that could scale to very large datasets.

2.3. Assessment and applications of design patterns

This sub-section discusses the CDS results obtained from the discussed literature related to the component model. Ma et al. [17] discovered 35 common design structures in a dataset of 120 B-rep models using a frequency threshold value of 4. The CDS computation time was 17.9 min with the 120 models having an average of 56.14 nodes in their FAGs. The researchers noted that their implementation required significant memory because of the need to store all candidate subgraphs while calculating their overall frequency in the dataset.

Bai et al. [4] described an algorithm for use with feature-based designs and tested the proposed approach using 438 CAD models. The algorithm extracted 36 CDSs from these models. An assessment of the results conducted with an expert in mechanical engineering found specific functions associated with most of the extracted design patterns. However, the expert could not be able to identify functions for some of the CDSs due to their complexity. It took less than 10s and 10m for clustering 1,500 and 10,000 models respectively, and less than 0.3s and less than half a minute to retrieve CDS from 280 models and 20,000 models respectively. They demonstrated that extracted CDSs could be reused in both door and bezel designs.

From the literature, the CDS assessment parameters used are the number of identified CDSs, retrieval time, and the effectiveness of reusing CDS application in design problems. The smaller number of CDSs identified in [17] and [4] approaches could be due to the larger number of constraints on CDSs during the extraction process. The time taken to extract CDS could be reduced with alterative part representation and data mining approaches.

The review of the graph-based approaches suggests that the trend is to add more geometric and topological information to the nodes and arcs of the network to generate structures with more semantic significance. The reported graph-based approaches for design structure retrieval focused mainly on topologically adjacent interactions between features and feature relationships. Such an approach can identify design structures on, or adjacent to, an individual face, for example, the gold or black faces on the valve body in Fig. 1. In contrast, the authors' approach will determine which features frequently occur in combination (e.g. Pitch circle diameters A & D & E or bores B & C) irrespective of their complexity or relative locations (e.g. adjacent or intersecting). Lastly, the results presented in the literature focused on the effective reuse of design structure in new product development. Whereas, our work primarily aims to analyse and understand how effectively a company can reuse CDSs by identifying substitutable features. The next section describes the proposed approach and details the scope of this research

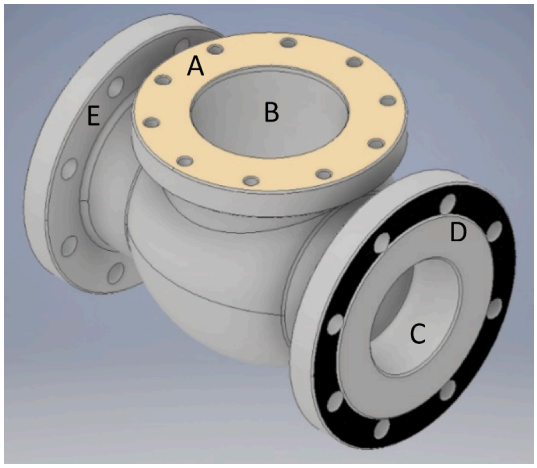


Fig. 1. Patterns of holes in a valve design.

work.

3. Research aim and the proposed approach

The aim of this work is to discover frequently occurring, common design structure (CDS) for hole features in an industrial valve dataset and illustrate the potential of the extracted CDSs by identifying substitutable features. In the extracted CDSs, the hole features do not have to be a constrained to specific locations or spatial connectivity between other hole features on a component. Fig. 2 illustrates the process of extracting hole CDSs on components. The first step involves extracting 3D features from CAD databases using a 'Twig Match' algorithm [18]. This algorithm uses a B-rep face adjacency graph representation and a sub-graph isomorphism matching process to search for 3D features (i.e. feature recognition using feature relationship graphs). The recognition system is very efficient and can search thousands of 3D models to find the required features in a matter of seconds.

In this paper, hole diameters were extracted from each component in the dataset. To extract hole diameters, the Twig match algorithm requires the B-rep of each of the components in the dataset which is generated using an open-source C++ 3D modelling library called Opencascade. Using this library, the B-rep was extracted from the components and structured into a graph, which was used by the Twig match algorithm to extract features. The vertices of the graph represent the faces of the component and edges represent the edges connecting the adjacent faces. The vertices have properties which includes the surface type, a face could be a plane, cylinder, cone etc, face convexity which could be flat, concave or convex, the face positional coordinates and if the face is cylindrical the radius can also be added as a property. All these properties are calculated using functions that are readily available in Opencascade.

The Twig match algorithm requires a definition of a hole which is defined as a cylindrical face joined by two flat planes, then the Twig match algorithm will match this definition and extract the sections of the B-rep that makes a hole feature, next the properties of the cylindrical face of the hole was used to extract information about the diameter and

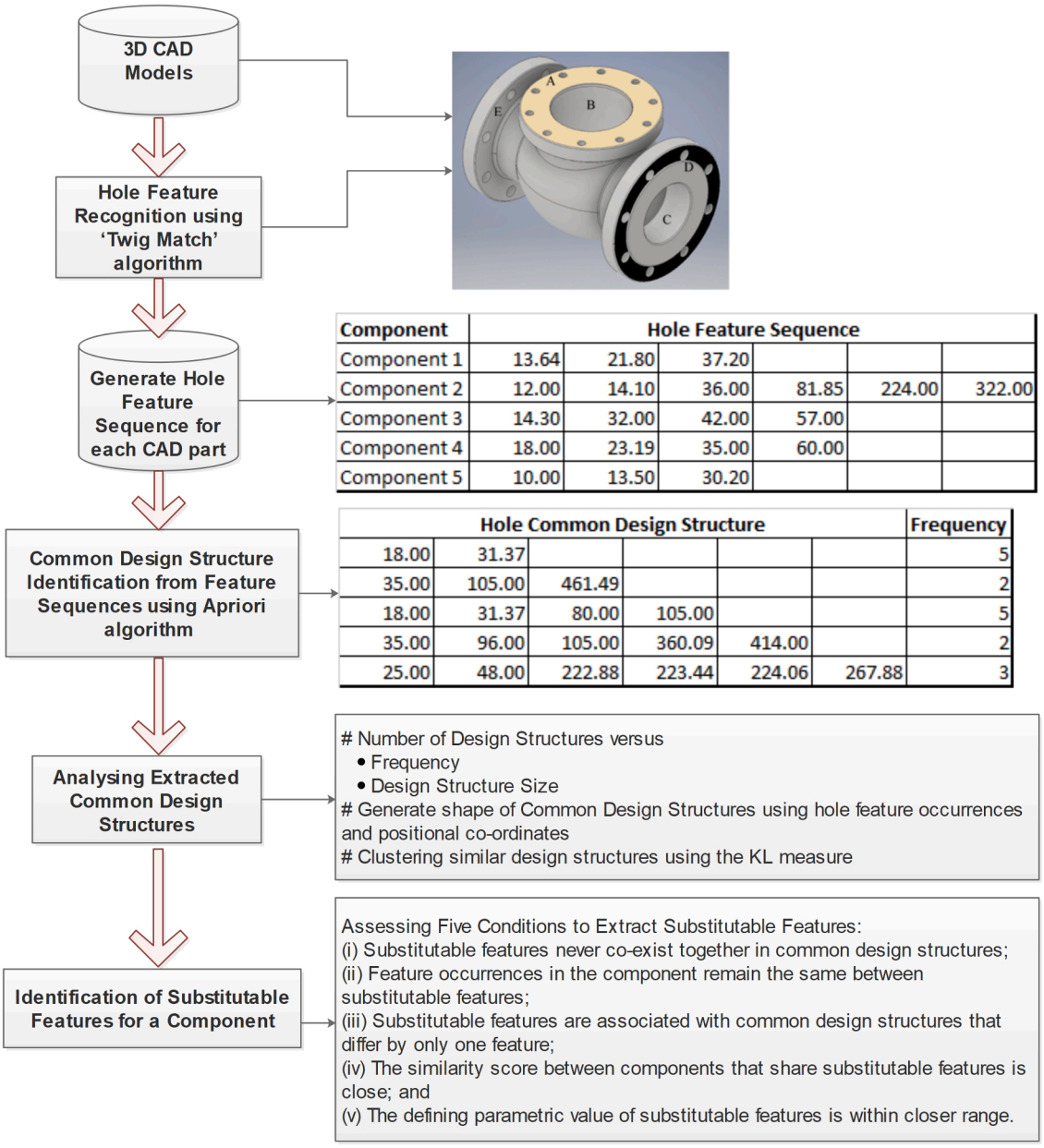


Fig. 2. Process for the discovery of CDS of hole feature and substitutable features.

the positional coordinates of the holes. The algorithm also extracted the information of the occurrences of hole diameters (i.e. the number of times a hole is present on a component), this was done after the hole diameters are extracted from the components and each of the hole diameters, in the list of extracted hole diameters, are tallied for their number of occurrences. In the next step, each component's lists of hole diameters were arranged as a sequence. Each sequence represents different hole diameters found in a component (i.e. no repetition of the same hole diameter size in a sequence), and the hole features were ordered in ascending order to facilitate the identification of frequently occurring CDSs.

The following step were involved in discovering the frequently occurring hole CDSs using the Association Rule Mining principle. The popularity of Association Rule Mining has spawned a large number of algorithms for identifying frequently occurring patterns in sequences of numbers (e.g. [9,11]). For illustration, we used the Apriori algorithm [1] to identify frequently occurring sets of hole diameters that are found together on components. This algorithm is one of the pioneering approaches to discovering frequent patterns in transaction databases. The pattern extraction algorithm requires what is termed a 'minimum support value' (i.e. the number of times a pattern occurred in the sequences) to be specified to define the limit of the set extraction process. This support value will vary based on the data sparsity in the respective datasets. The support value will be low if the sparsity is high, and vice versa. Since decreasing the support value would increase the algorithm's running time, a support value needs to be selected which will recover a reasonable number of valuable patterns within an acceptable run time. This step produces a list of frequently occurring hole feature sets along with the frequency information.

We have used the SPMF open-source data mining library [10] to extract the CDSs. The subsequent step analysed the generated hole CDSs using the size (i.e. the number of hole features in a structure), and frequency of occurrences in the dataset. The analyses also involves the shape of hole design structures generated using hole feature occurrences and positional coordinates. The shape of the hole design pattern represents where the holes are in the component. Fig. 3 illustrates the shape of a hole CDS generated (25.4 and 254.0 hole diameters) in a component. The shape of hole design structures along with the design structure size and frequency of occurrences will help to understand the distribution of hole design structures in the dataset. The potential of extracted hole CDSs was demonstrated in identifying substitutable features in the CAD dataset. A Kullback–Leibler (KL) divergence measure is proposed to identify the similarity between components that share common CDSs. This measure will be useful in finding substitutable features as one of the conditions is that the similarity score between components that share CDS has to be close. The following sections illustrate the proposed approach with a dataset of industrial valves described below.

4. Dataset description

A valve design dataset was created from an online catalogue of industrial components. In total 1851 3D models of the industrial valve were downloaded from several manufacturers. Using the Twig Match algorithm, a hole feature sequence was generated for each CAD model by extracting hole features in 3D CAD models. The hole feature sequence contains hole features with different diameters. A hole feature is not allowed to appear twice in the same sequence used to generate CDSs across hole features. For example, the sequence of hole diameters associated with a component shown in Fig. 3 is {25.4, 54.56, 254.0}. The weblink for the extracted dataset of hole features from the valves models is provided in the Acknowledgements section. In total the dataset contains 796 different hole diameters. Fig. 4 shows the frequency count of the number of different hole diameters in a component. Components with distinct one- and two-hole diameter features cover 48.5% of the dataset, and three components have a maximum of 11 different holes. Fig. 5 shows that 650 (82.4%) of the different hole diameters occur only between 1 and 7 times across components in the dataset. The low frequency of use illustrates the sparsity of hole diameter reuse in the dataset, which will greatly influence the CDS extraction. The following section details the CDSs extracted from hole feature sequences and the identification of substitutable hole features.

5. Results

5.1. Common design structure description

The Apriori algorithm [1] was used to extract frequently occurring

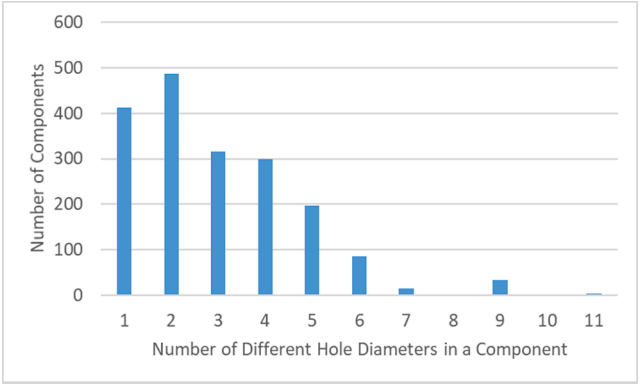


Fig. 4. Frequency of number of different hole diameters in a component.

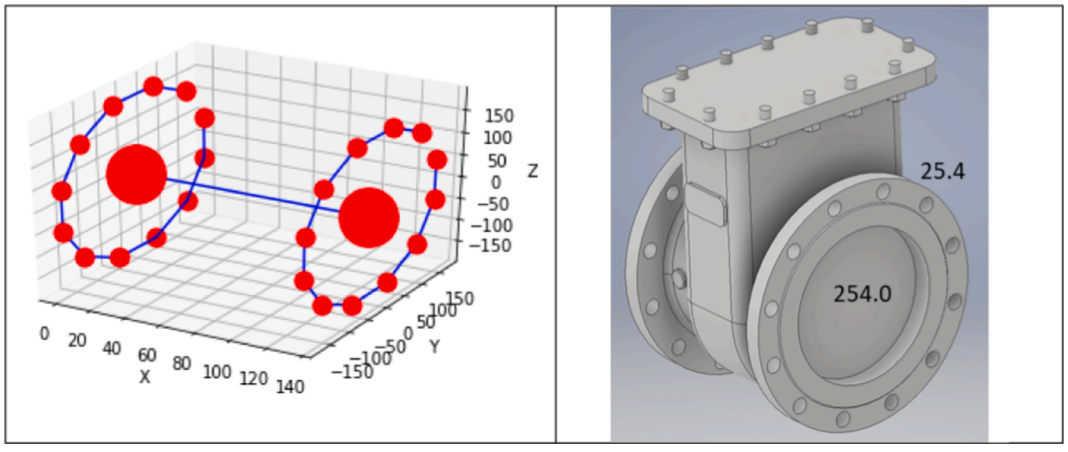


Fig. 3. The shape of a ([25.4, 254.0]) generated hole design structure and its associated 3D component.

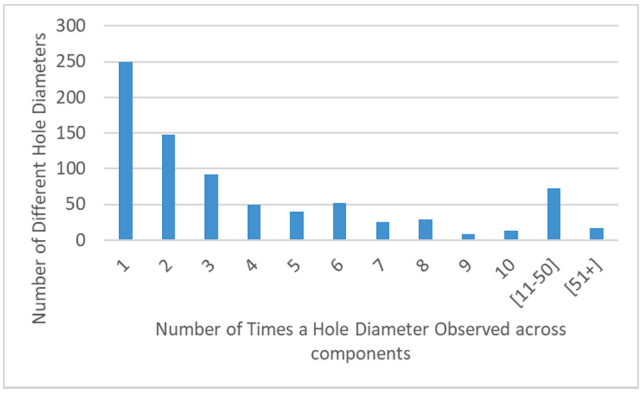


Fig. 5. Distribution of hole presence across designs.

CDSs from the hole diameter feature sequences that were generated from each 3D CAD component. Fig. 6 details the number of generated hole CDSs and run time for each level of support percentage. The support of a CDS is how many times the CDS appears in the hole sequences. The time taken, although in milliseconds, has significantly increased with 0.1 support percentage (i.e. the CDSs occurred at least twice in the generated sequences) but generated a maximum number of CDSs from the dataset. This support percentage value is low because the dataset contains sparse hole diameters spread across many sequences, as highlighted in the distribution of holes' presence across components (Fig. 5). Further analysis uses the hole CDSs generated at the support percentage of 0.1%. The algorithm has taken 2177 ms to generate 8454 frequent CDSs at 0.1% support value, which includes CDS size varying from 1 to 9 different hole diameters. The CDS size represents the number of hole diameters in a CDS. The number of CDSs reduces with the increase in the CDS size (Fig. 7). 80% of the CDSs contain between two and five unique hole diameters.

Fig. 8 depicts the reuse frequency of common hole design structure in the valve dataset. The figure shows that nearly 58% of the CDS occurred only twice in the dataset. However, a single hole pattern has been identified for a maximum of 204 times. Out of 1851-hole sequences associated with components in the dataset CDSs are generated for 1832 sequences (99%). Interestingly 1% of the sequences have 33 different hole diameters that have never been reused in other CAD models. The distribution graph in Fig. 9 illustrates that 51% of the components contain less than four CDSs. However, the maximum number of CDSs is up to 511 for 18 components. The maximum 511 design patterns occur when nine different hole diameters occur on a component. Each component could generate a maximum number of CDSs of $2^n - 1$, where n is the number of different hole diameters. But many of these CDSs do not occur as they do not pass the support threshold and many

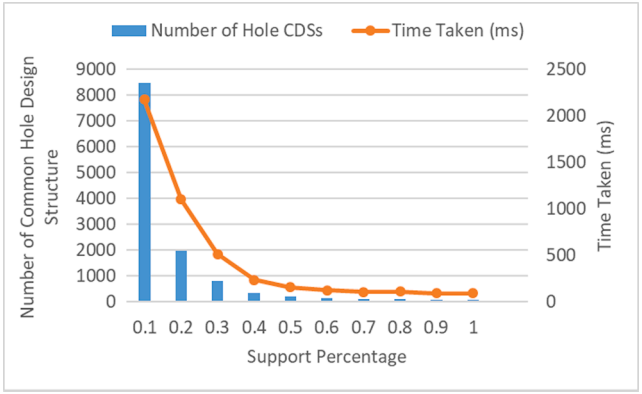


Fig. 6. Number of common design structure and time taken with reference to the support percentage.

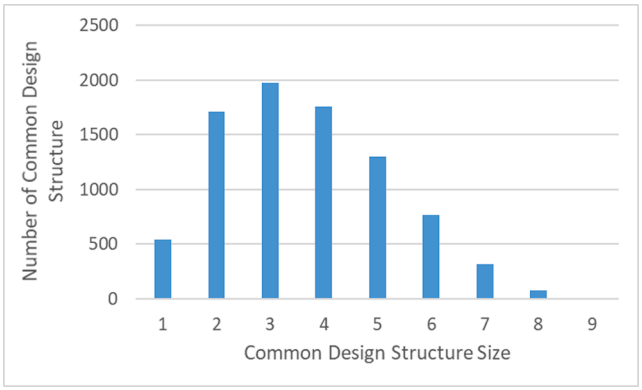


Fig. 7. Distribution of number of common design structure to design structure size.

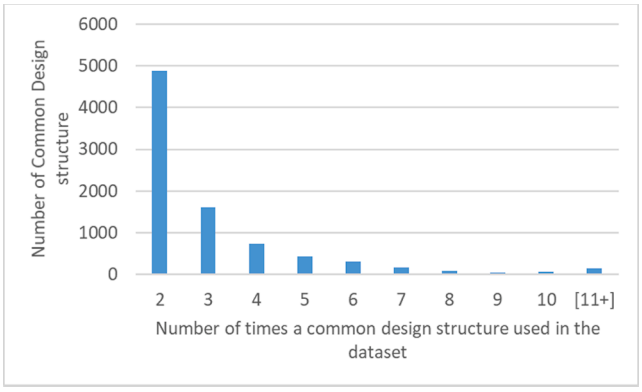


Fig. 8. Reuse frequency of design structures in the valve datasets.

components often still use different hole diameters. The trends that are shown in these graphs will be helpful in the next section that aims to find applications for the extracted hole CDSs.

5.2. Shapes of most frequently occurring hole common design structure

The shape of common hole design structures is an important element in analysing and discovering emerging patterns. The shape analyses play a vital role in effective understanding and subsequently reusing the CDS in new designs. The X, Y, and Z coordinates extracted from the hole locations were used to construct the shape for each hole CDS. In the dataset, the most frequently occurred CDS is an 18mm hole diameter. The 18mm hole diameter has been used in 204 components. Table 2 summarises the various frequently occurring CDS shapes for the 18mm hole diameter across the 204 components. The shape analysis of 18mm hole diameter reveals the applications of this hole diameter across four different component types, the different shape structure within each component type, and the number of components shared the same CDS. These discoveries are useful in identifying the application of a feature across component types. In this case, the 18mm hole diameter is commonly used for bolt connection.

5.3. K.L. measure for similarity of components

Although the common design structures are shared across components, there will be variation within these components. As such, we can view the location of the hole coordinates as having been realised from a multivariate probability distribution and therefore we can use the actual locations as data to estimate the distribution. This will result in one multivariate distribution for each component. The Kullback–Leibler (K.

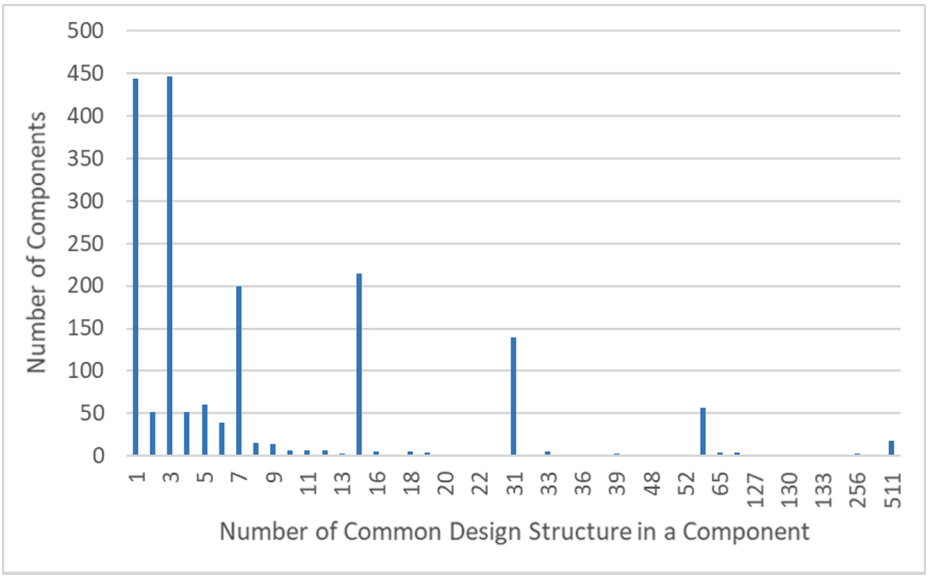


Fig. 9. Distribution of number of common design structures in components.

L.) divergence provides a means to measure the difference between probability distributions [15]. Using the K.L. divergence measure in this way to measure the difference between components is superior to a Euclidean distance measure as it does not require the CDSs to contain the same number of coordinate points. The K.L. divergence score of 0 indicates that the hole positional coordinates between two components are identical and the higher the measure implies higher variation between two components. Considering Q and P are the two components, the formulation of K.L. measure is detailed as follows.

The K.L. divergence from Q to P is:

$$D_{KL}(P||Q) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(x,y,z) ln\left(\frac{p(x,y,z)}{q(x,y,z)}\right) dxdydz$$

where p(x,y,z) and q(x,y,z) are the probability density functions of P and Q describing the likelihood of a hole being located at coordinates (x,y,z).

The K.L. is a measure of divergence, not distance and as such $D_{KL}(P\|Q) \neq D_{KL}(Q\|P)$.

To estimate the multivariate probability density function (pdf) we use a Kernel density estimate which has the following form:

$$p(x,y,z) = \frac{\sum_{i=1}^{n_p} \phi\left(\frac{x-x_i}{\sigma}\right)\phi\left(\frac{y-y_i}{\sigma}\right)\phi\left(\frac{z-z_i}{\sigma}\right)}{n_p}$$

where:

$\phi()$ is the standard Normal distribution pdf

n_p is the number of holes in design P

(x_i, y_i, z_i) is the coordinates of the i^{th} hole

Kernel density estimation [2] is a non-parametric approach to probability density estimation, using probability density functions centred on each datum to support data smoothing. The result is a pdf that is defined for the whole range of coordinates not just where there are data. The smoothing parameter or bandwidth for this model is σ, which controls how quickly the pdf will decrease as it moves away from an observed coordinate. We are free to choose different bandwidths for each coordinate but for simplicity have chosen to keep them the same.

Substituting the Kernel density estimates into the K.L. formula produces the following.

$$D_{KL}(P||Q) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \left[\frac{\sum_{i=1}^{n_p} \phi\left(\frac{x-x_i}{\sigma}\right)\phi\left(\frac{y-y_i}{\sigma}\right)\phi\left(\frac{z-z_i}{\sigma}\right)}{n_p}\right] ln\left(\frac{\frac{\sum_{i=1}^{n_p} \phi\left(\frac{x-x_i}{\sigma}\right)\phi\left(\frac{y-y_i}{\sigma}\right)\phi\left(\frac{z-z_i}{\sigma}\right)}{n_p}}{\frac{\sum_{j=1}^{n_q} \phi\left(\frac{x-x_j}{\sigma}\right)\phi\left(\frac{y-y_j}{\sigma}\right)\phi\left(\frac{z-z_j}{\sigma}\right)}{n_q}}\right) dxdydz$$

This will require numerical methods to calculate. We can simulate, i.e. Monte Carlo, as we can express this as an expectation with respect to the probability measure for$P$.

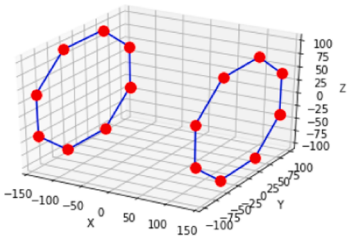
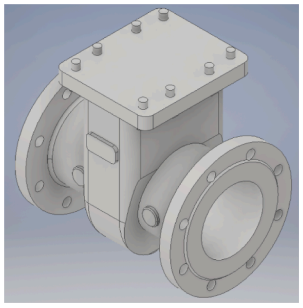
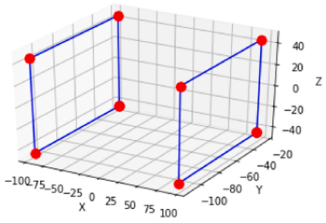
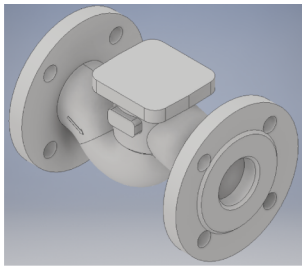
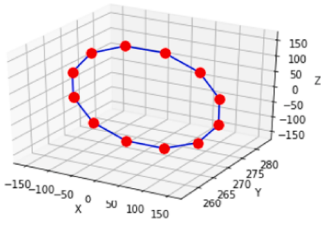
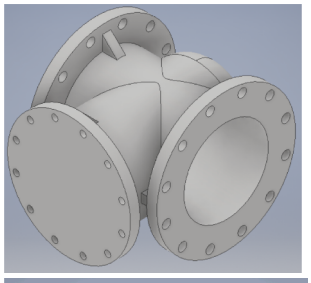
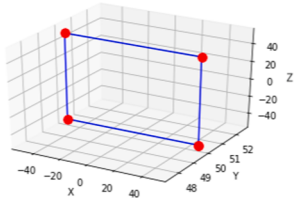
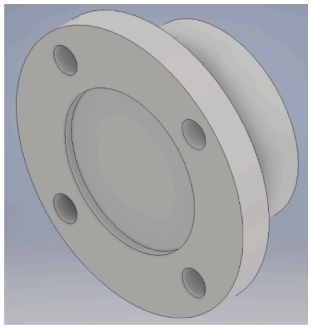
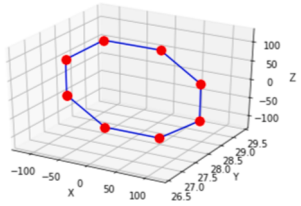
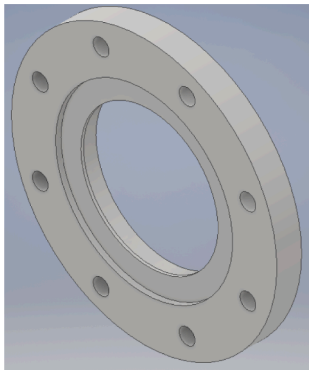
$$D_{KL}(P||Q) = E\left[ln\left(\frac{\sum_{i=1}^{n_p} \phi\left(\frac{x-x_i}{\sigma}\right)\phi\left(\frac{y-y_i}{\sigma}\right)\phi\left(\frac{z-z_i}{\sigma}\right)}{n_p}\right)\right] - E\left[ln\left(\frac{\sum_{j=1}^{n_q} \phi\left(\frac{x-x_j}{\sigma}\right)\phi\left(\frac{y-y_j}{\sigma}\right)\phi\left(\frac{z-z_j}{\sigma}\right)}{n_q}\right)\right]$$

Simply through Monte Carlo simulation from density $p(x,y,z)$ we can evaluate the average value of the above K.L. measure.

5.4. Clustering hole common design structure

A common design structure (**10.0, 19.05, 32**) that contains three-hole diameters was used to illustrate the clustering process using the K.L. measure. The CDS (**10.0, 19.05, 32**) was shared across 11 components. Table 3 shows the two different shapes identified in the common design structure of (**10.0, 19.05, 32**). Table 4 provides the comparative K.L. measure score between these 11 components. The lowest (i.e. similar components) and highest (i.e. dissimilar components) K.L. scores are highlighted in the table. The Hierarchical

Table 2
Shapes of a common design structure for 18 mm hole diameter.

Valve component types	The shape of 18 mm Hole Common Design Structure	Actual component	Number of components shared this same CDS
Body			44
			49
			15
Flange			28
			17
Stopper			40

(*continued on next page*)

Table 2 (*continued*)

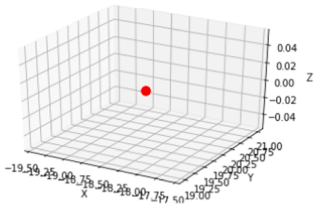
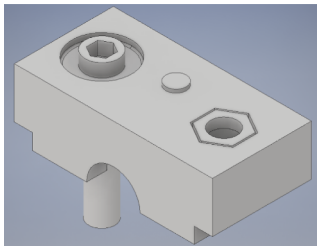
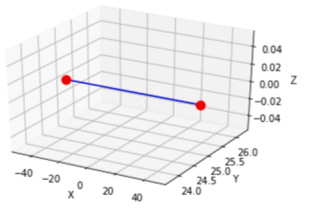
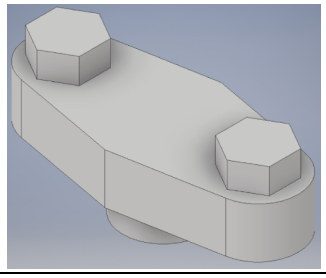
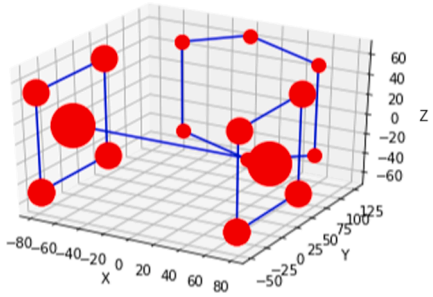
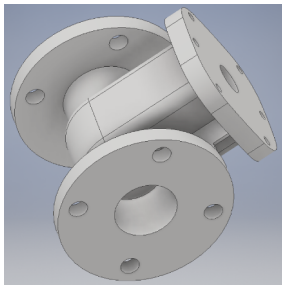
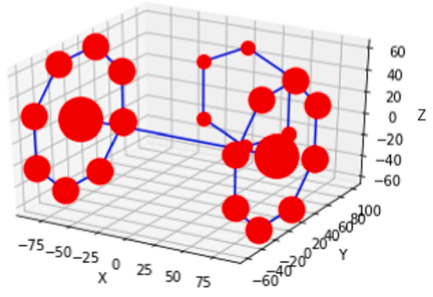
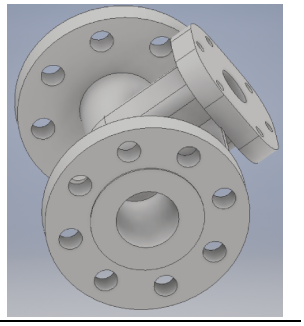
Valve component types	The shape of 18 mm Hole Common Design Structure	Actual component	Number of components shared this same CDS
Gland			11
			

Table 3
Shapes of common design structure ([10.0, 19.05, 32]).

Shape of {10.0, 19.05, 32} Hole Diameters Common Design Structure	Actual component
	
	

clustering process was used to create the similarity clusters. Since the K. L. measure for two components is not commutative (i.e. $D_{KL}(P\|Q) \neq D_{KL}(Q\|P)$), the combined absolute scores are used for the clustering process. A symmetric distance matrix with these combined scores was generated to support the clustering process. The shortest distance in the matrix using the complete link approach was used to generate the clusters. Fig. 10 represents the generated dendrogram for the complete link cluster. The clustering order was accurately matched with the conducted manual assessment (Fig. 11). The clusters grouped similar components, and the variations between them are highlighted with the link values in the dendrogram diagram. The K.L. measure will be useful in the process of finding substitutable features, which is illustrated in the next section.

5.5. Identification of substitutable hole features

An important application of common design structures, illustrated in this paper, is to find substitutable hole features. The common design structure represents features that are frequently occurring together in various components. From these common design structures, this paper illustrates an approach that elicits substitutable features. Substitutable features are interchangeable between components that have the same function without changing the significance of the structural appearance. Table 5 lists the conditions for substitutable features and the rationale for those conditions.

Two approaches can be enabled for identifying the substitutable features based on the needs of engineers. In the first approach, the engineer can choose a component and look for a possible substitutable

Table 4
K.L. measure scores between 11 components that share a common design structure ([10.0, 19.05, 32)] (Lowest and highest scores are highlighted in different grey shades).

P\Q	1	2	3	4	5	6	7	8	9	10	11
1	0	0.16	0.03	0.54	0.54	0.13	0.59	0.13	0.56	0	0.78
2	0.03	0	0.03	0.76	0.76	0	0.79	0	0.75	0.03	0.57
3	0.01	0.13	0	0.52	0.52	0.15	0.55	0.15	0.54	0.01	0.82
4	0.14	0.35	0.13	0	0	0.37	0.05	0.37	0.09	0.14	1.15
5	0.15	0.35	0.13	0	0	0.38	0.05	0.38	0.09	0.15	1.16
6	0.03	0.5	0.01	0.77	0.77	0	0.83	0	0.77	0.03	0.57
7	0.16	0.29	0.09	0.02	0.02	0.39	0	0.39	0.07	0.16	1.19
8	0.03	0.05	0.01	0.79	0.79	0	0.85	0	0.79	0.03	0.55
9	0.13	0.26	0.13	0.22	0.22	0.28	0.25	0.28	0	0.13	0.72
10	0	0.17	0.04	0.55	0.55	0.13	0.61	0.13	0.57	0	0.77
11	0.78	0.71	0.81	1.91	1.91	0.67	1.95	0.67	1.15	0.78	0

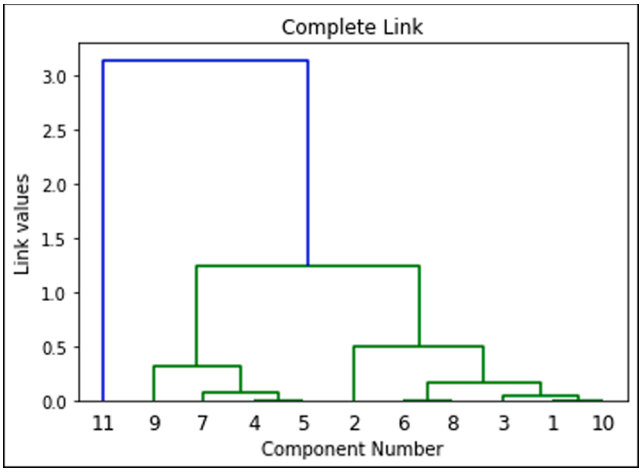


Fig. 10. Dendrogram of the complete link cluster.

feature within the component. In the second approach, the engineer can browse through all the substitutable features from a knowledge-based system. The procedure for extracting substitutable features for a chosen component is explained in a step-by-step approach in Table 6 (and can be easily modified to identify all the substitutable features in the dataset). The procedure for identifying "substitutable" features that satisfy the conditions mentioned in Table 5 from the extracted common hole design structures was coded in Python.

A component with four-hole features is used to illustrate the first approach. Fig. 13 portrays the selected component. The four-hole diameters are {10, 19.05, 32, 63.5}. The results derived from the step-by-step procedure illustrated in Table 6 are subsequently elaborated. The selected component has 15 common design structures. It means that all combinations of hole diameters are frequently occurring in the dataset (i.e. $2^n - 1$ combination). The selected component matched with 85 components that have at least one common design structure. Among 85 components, 82 components have only one common design structure shared with the selected component. The maximum number of shared common design structures is seven. In total, 1,322 different CDSs were identified from these 86 components (including the selected component). The 1,322 CDSs were filtered to 587 CDSs that contain at least one-hole diameter of the selected component. Comparing these 587 CDSs with each other produces 6632 pairs of hole diameters where CDSs differ by a single one-hole feature.

In these hole pairs, 2802-hole pairs were filtered that never co-occurred in the dataset. The cut-off value for finding possible substitutable hole diameters for the chosen four-hole component was fixed at 30mm. This cut-off value further reduces the possible substitutable hole

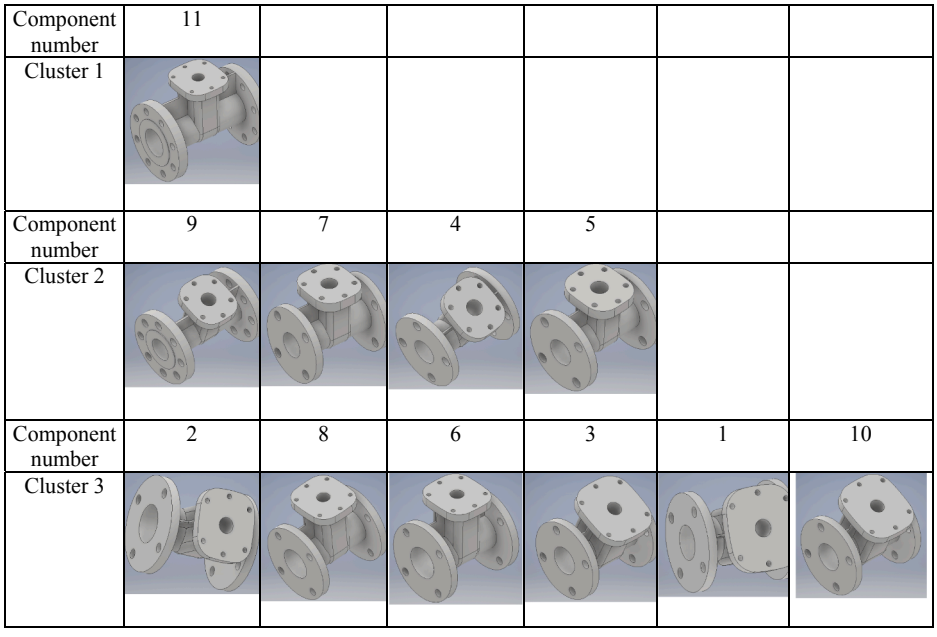


Fig. 11. Hierarchical clustering of valve body components that share a common design structure ([10.0, 19.05, 32]) using the K.L. measure scores.

Table 5
Conditions for substitutable features and their rationale.

Conditions for substitutable features	Rationale
Substitutable features never co-exist together in common design structures.	Substitutable words never co-occur in a sentence. The same analogy is applied to CAD models.
Feature occurrences in the component remain the same between substitutable features.	The same number of times substitutable features occur in components will ensure the significance of the structural appearance. The location pattern of the substitutable feature remains the same.
Two common design structures have a one-hole feature difference between them.	The concept of finding one-hole feature difference between CDSs is based on the concept of Triadic closure in network analysis. Triadic closure defines a common component that shares features with two separate components. Fig. 12 illustrates the triadic concept with two observed CDSs. The two CDSs (10, 19.05, 32, 63.5) and (10, 19.05, 32, 76.2) have three-hole diameters in common, and the hole diameters 63.5 and 76.2 are different. This one-feature difference between CDSs has the potential substitutable opportunity.
The similarity score between components that share substitutable features is close.	Restricting the difference in the similarity score will ensure the substitutable features belong to the same component type since the similarity score is based on the feature coordinates in the component.
The defining parametric value of substitutable features is within close range.	This condition is based on the logic that the substitutable features will be within a close range of parametric values. The difference of 30 mm between hole diameters is chosen for this study.

pairs to 301, where duplicate hole pairs were also removed. From these 301-hole pairs, 45-hole pairs were selected that contain any of the hole diameters of the selected component. No possible substitutable holes were identified for 10 mm hole diameter. Table 7 summarises the possible substitutable holes for the other three-hole diameters. The possible substitutable holes mentioned do not co-occurred with the holes in the selected component in any other components and the variation is less than 30 mm.

The next two steps in the process are to identify the occurrence match between the hole diameter in the selected component and the components that contain possible substitutable hole diameters, and check the substitutable suitability through component similarity scores. Table 8 summarises the occurrence match between the hole diameter in the selected component and the number of components which matched the occurrence of hole diameter in the selected component, that contain possible substitutable hole diameters. Some possible substitutable hole diameters do not match the occurrence of hole diameter in the selected component. These substitutable hole diameters were removed for the final step to calculate a component similarity score for the group of components that matched the hole diameter occurrence. The removed hole diameters are highlighted in red colour in Table 7. Table 8 summarises the number of components containing possible substitutable hole diameters that matched the occurrence of hole diameter in the selected component. The final step involves calculating the K.L. measure to identify the similarity between the selected component and these matched components.

Using the proposed K.L. measure the similarities between the selected component and the components that contain possible substitutable hole features were assessed. The K.L. measure of zero represents a high likelihood of using the substitutable hole in the selected component because there is a greater similarity between components. Table 9 lists the best K.L. measure score (i.e. the minimum score) and the

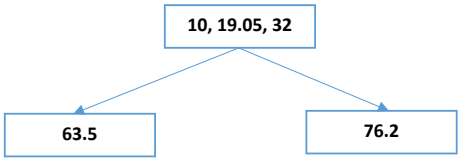


Fig. 12. Illustration of Triadic closure concept to identify substitutable features in two CDSs ([10, 19.05, 32, 63.5]) and ([10, 19.05, 32, 76.2]).

Table 6
The pseudo-codes to extract substitutable features for a chosen component.

Input:
Selected component
Output:
Substitutable hole features for the selected component
Algorithm:
Step 1: Get all common design structure for the selected component
Step 2: Get all components that have at least one CDS common to the selected component
Step 3: Get all CDSs that contain at least one-hole diameter of the selected component and compile a CDS list with no duplicates
Step 4: Loop to compare each CDS with every other CDSs in the list
Step 4a: Identify one-hole feature difference between the compared CDSs and form a pair with two-hole features that differ between two CDSs
Step 4b: Check if the hole pair contains a hole feature from the selected component
Step 4c: Check if the hole pair not co-occurred in the complete hole sequences
Step 4d: Check the difference between hole diameters is less than 30 mm
Step 4e: Group the hole features into a set that has a common connection but not co-occurred (ignoring any repetitive hole features). The number of sets generated will be equal to the number of hole features in the selected component.
Step 5: Loop to compare hole features within each generated set
Step 5a: Select the hole diameters that match occurrences of hole feature in the selected component.
Step 6: Loop to compare the hole features within each occurrence-matched generated set
Step 6a: Calculate the K.L. measure between the selected component and the component that contain the hole feature in the occurrence-matched hole-feature set.
Step 6b: Order the similarity score in ascending order with reference to the hole feature. The top-most-hole diameters are most likely to be the substitutable hole features for the selected component.

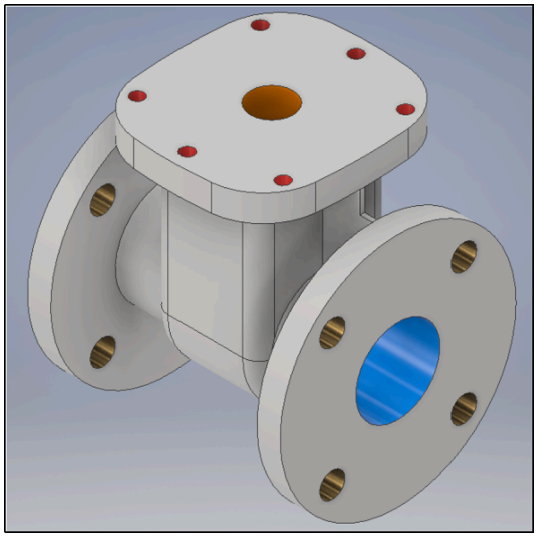


Fig. 13. The chosen four-hole feature valve body used to illustrate the process of identifying substitutable features (CDSs associated with hole features are highlighted in different colours). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

CAD image of a component that contains the substitutable hole diameter feature for 63.5 mm hole diameter. The analysis of the K.L. score and the associated CAD image support the following observations:

- Eight out of 13 possible substitutable hole diameters (that are listed in Table 7) were found to be useful for substituting 63.5 mm hole diameter.
- Eight identified substitutable hole diameters are valid as all these diameters represent bore diameter in the valve body and share similar topological structure.
- It can be observed that the primary structural variation linking to the presence of hole diameters between the selected component and the component that contain substitutable feature increases with the K.L. measure. The K.L. score above of four represents the largest variation with reference to the selected component, and is adopted as a cut-off (i.e. threshold) score to eliminate the substitutable hole diameters.

The five conditions established in this process have resulted in the accurate identification of substitutable features. The substitutable results obtained for the other two-hole diameters (19.05 and 32 mm) are illustrated in Appendix – 1, which again show the high accuracy of the identified substitutable hole features generated by this approach. Also, for 19.05 mm, the K.L. scores of above four identifies it as not being substitutable features. Except for a substitutable feature for 32 mm, all other identified substitutable holes features are correct. The only incorrectly identified substitutable feature of 50mm hole diameter (where the K.L. score is less than four) for 32 mm stem diameter was not correct because it represents bore diameter. Table 10 details the final list of substitutable holes for the selected component in the order of priority.

6. Discussion

The paper has demonstrated how the mining of frequent set of holes in mechanical components can be used to identify common design structure in large CAD datasets. Although the hole's shape is represented by a single number (its diameter), the resulting design patterns provide sufficient characterisation to provide insights into the structure of the dataset. Also, irrespective of any spatial connectivity between other features on a component, the extracted design structures find application in identifying substitutable features. The important point emphasised in this work is that all CAD features have the potential for reuse irrespective of their connectedness, dependency and complexity. The proposed approach is not limited to holes and the system can easily be extended to other feature types whose form can be defined parametrically.

The itemset-based approach to identify common design structure leads to the use of an efficient data mining algorithm that is computationally efficient. This approach provides an alternative approach to the feature- or face-graphs that are commonly used to find CDS. No labelling of features or qualitative annotations is required for this approach. This unlabelled approach has benefits that avoid misinterpretation between engineers due to purely parametric description. Also, the approach does not require any classification of components. The approach helps to generate many CDSs (8454 CDSs @ 0.1% threshold frequency) in less time compared to the results reported in the literature. The importance given only to the frequency in the extraction process would be more valuable to find many applications using CDSs. The visualisation of CDSs with hole feature occurrences and positional coordinates in a graphical format helped to understand the difference of hole design structures within and across CDS in the dataset. The graphical representation also illustrates the richness of information in geometrical topology structure within hole features of different component types.

The number of CDSs generated is considerably less even for 0.2% support value, which demonstrated the sparsity of feature reuse in this dataset. The CDS provides better representation if it contains more features, but the maximum number of hole CDSs is for the size of three features. Improving feature reuse could enable to increase the CDS size, which will subsequently improve the expressiveness content in it.

Table 8
Number of components containing possible substitutable hole diameters that matched the occurrence of hole diameter in the selected component.

Hole diameter in the selected component	The occurrence of hole diameter in the selected component	Number of components containing possible substitutable hole diameters that matched the occurrence of hole diameter in the selected component
19.05	8	191
32	1	74
63.5	2	91

Table 7
Possible substitutable hole features for the hole diameters in the selected component (the component numbers highlighted in red colour do not match the occurrence of hole diameter in the selected component). (For interpretation of the references to colour in this table text, the reader is referred to the web version of this article.)

Hole diameter in the selected component	Possible substitutable hole diameters	Number of possible substitutable hole diameters
19.05	10.11, 11, 12, 13.5, 14, 17.5, 18, 19, 22, 22.23, 25, 28, 30, 30.2, 40, 42, 45, 48	8
32	10.11, 11, 12, 13.5, 14, 17.5, 22, 27.28, 30.2, 45, 50, 51	7
63.5	34, 40, 42, 45, 48, 50, 50.8, 51, 55, 57, 65, 68, 70, 76, 76.2	13

Table 9
Best K.L. measure score and the CAD image of a component that contains the substitutable hole diameter feature for 63.5 mm hole diameter (hole diameters that are identified as not suitable for substituting 63.5 mm diameter based on higher K.L. score are highlighted in red colour). (For interpretation of the references to colour in this table legend, the reader is referred to the web version of this article.)

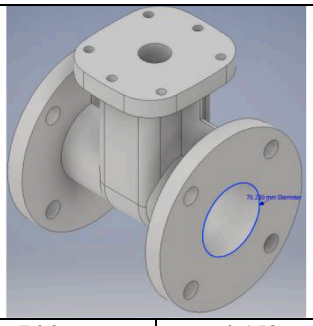
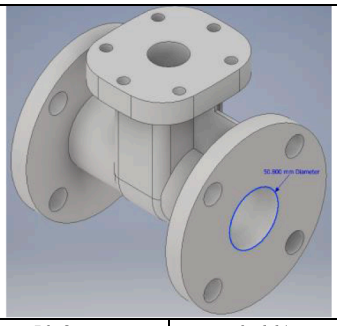
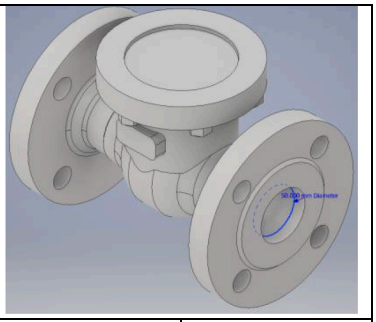
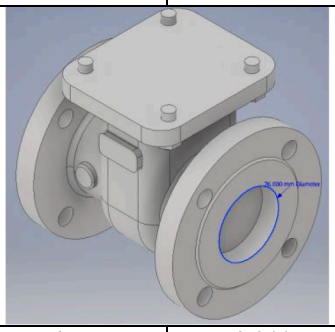
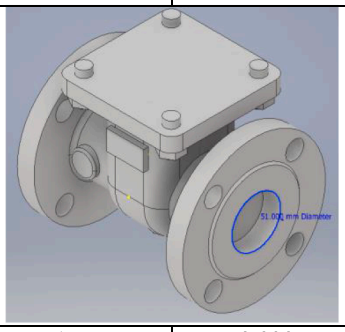
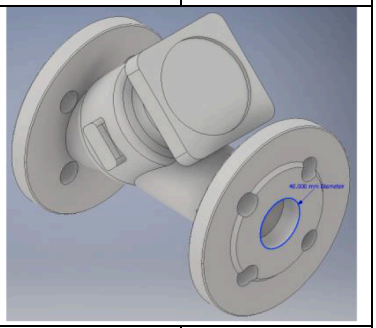
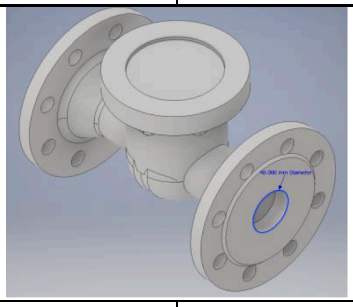
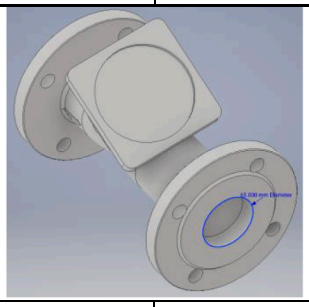
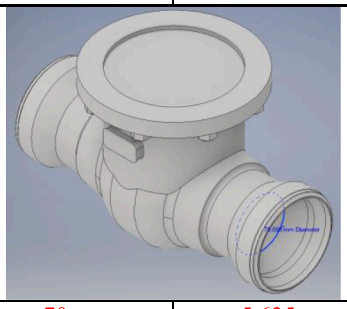
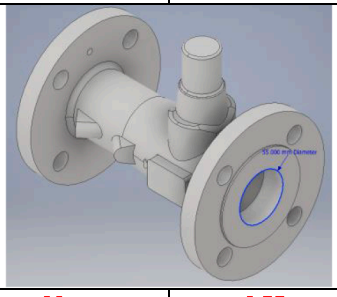
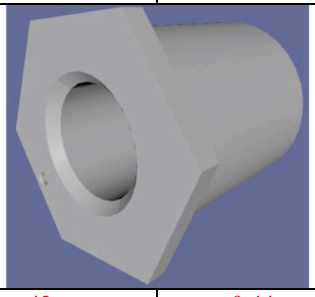
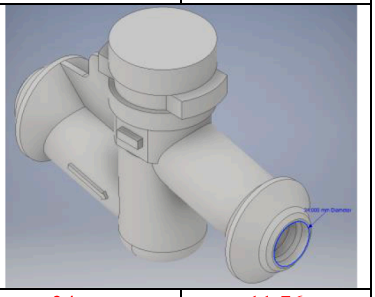
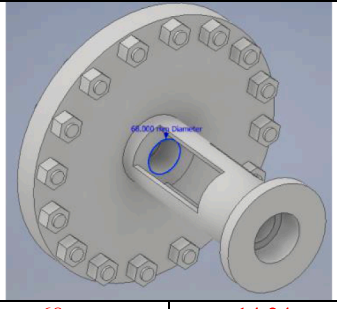
					
76.2 mm	0.158	50.8 mm	0.661	50 mm	1.957
					
76 mm	2.044	51 mm	2.939	40 mm	2.959
					
48 mm	3.08	65 mm	3.55	70 mm	5.635
					
55 mm	5.77	42 mm	9.44	34 mm	11.76
					
68 mm	14.24				

Table 10
Final list of substitutable holes for the selected component in the order of priority.

Selected hole diameter	Substitutable holes
19.05	19, 18, 22, 14
63.5	76.2, 50.8, 50, 76, 51, 40, 48, 65

Although most of the components contain only one CDS, there are components with 9-hole diameters that generated CDSs of all 511 combinations. The presented frequency CDS distribution graphs will be useful to industries in both understanding the effectiveness of feature reuse and highlight opportunities to improve reuse by identifying substitutable features.

Because of the resources applied to the development of frequent itemset mining, extremely efficient implementations exist that are fast enough to support interactive applications. For example, Fig. 14 shows the interface of a prototype shape browser that uses the similarity between sets of frequently occurring holes to determine the relative location of components in the display. The common design structure browser will be helpful to the user for browsing over the design instances of interesting design structures related to the design solutions.

The proposed Kullback–Leibler divergence measure for comparing the similarity between components based on hole coordinates seems appropriate for clustering similar components and provides a criteria for a condition check for filter the substitutable features. The merit of K.L. measure is that it facilitates comparing two components, even if the number of hole coordinate points between them differs. The substitutable results illustrate that as the K.L. measure increases the suitability of the substitutable features decreases.

The following five conditions proved effective in identification of substitutable features:

- Substitutable features never co-exist together in common design structures.
- Feature occurrences in the component remain the same between substitutable features.
- Two common design structures have a one-hole feature difference between them.
- The similarity score between components that share substitutable features is close.

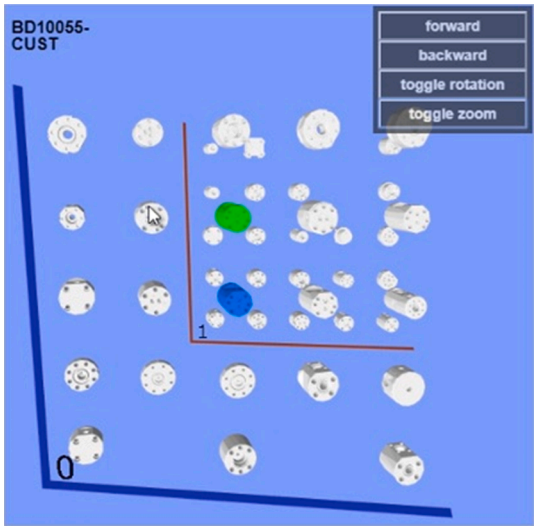


Fig. 14. Prototype interactive shape browser that dynamically arranges the shapes with similar sets of holes (i.e. design structures) closest to the blue query shape. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

- The defining parametric value of substitutable features is within a threshold value. Although the difference of 30mm between hole diameters has been chosen for this study, the results demonstrate that the difference of 20mm would be acceptable for this dataset.

Manual assessment of the accuracy of the substitutable features identified, using the five conditions, was high. Out of 13 identified substitutable features for the three selected hole diameters, only one substitutable feature was found to be incorrect. However, if the accuracy is measured only in terms of the substitutable features, that have the same function, the overall structural variability is reduced. This measure of accuracy assesses if the necessary modification to change between substitutable features are feasible. Although a feature's function is not explicitly represented in the dataset, the authors would argue that common function is implicit in the similar geometry of substitutable features identified. Although the accuracy of discovered substitutable features through the filtering process is apparent, characterising it using precision and recall curves is not appropriate because the number of "correct" substitutable features is a subjective measure dependent on the definition used. However, in the future comparative (i.e. relative) measures of the number of CDS and substitutable features identified by different algorithms could be used to judge performance. To enable different algorithms to test the robustness of the discovered features in this study (and if more substitutable features could be identified) the URL of the dataset is provided in the Acknowledgment section.

An advantage of the proposed approach is that since feature parameters were extracted from STEP files, the proposed extraction of CDS will work on CAD models from different proprietary CAD platforms. It should be noted that the substitutable features in this work are identified without defining the relationships between hole features such as dependency, intersecting and adjacency. However, nothing in the system's architecture precludes adding such feature relationships to the common design structures and such a development could potentially enhance the understanding of design patterns.

Although the system is described and implemented in terms of hole features, the approach could be expanded to more complex features. Indeed, in principle the algorithm described could be used with any form of parametrically defined features. One important constraint, however, is the need to identify the range of dimension values associated with substitutable features. The case study results identify range value of +/- 20 mm between substitutable values as being appropriate. However, this range value will vary with different types of engineering product and their associated industrial CAD dataset. So, although in practice this 'range value' could be based on the judgement of an expert engineer, it is also possible that an optimisation algorithm could be developed to enable the automation of the process. Another limitation of the prototype implementation is that it gives greater importance (i.e. high rank) to CDSs and substitutable features with similar sizes because it is assumed that a large change to feature size could lead to expensive manufacturing process alterations. However, this is not an inherent limitation and if the objective of the design reuse is to save, say, material or energy consumption during manufacture it is possible to add additional forms of filters to reflect these priorities.

Fig. 15 illustrates schematically how the CDS and substitutable feature discovery system could be integrated with parametric, feature-based, design modelling software. Central to such a system is the development of the CDS and substitutable feature library (database) from the automatic analysis of existing industrial CAD models. Given that well established algorithms are available for both efficient feature parameter extraction [18] and also the discovery of itemsets from the feature data of multiple components [11], there are no inherent computational issues in the identification of common design structure within large and complex CAD dataset. However, calculating the K.L. score (to identify the similarity between CAD models for detecting substitutable features) cannot be done at interactive speeds on a large CAD dataset. Consequently, the process of calculating the K.L. score

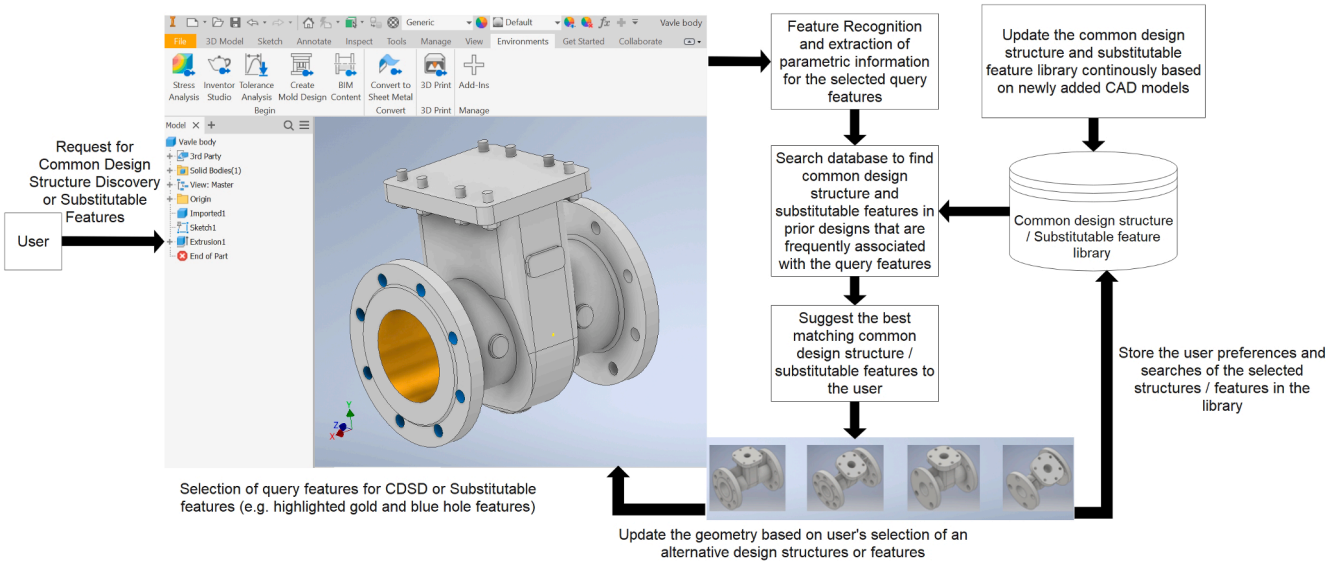


Fig. 15. Schematic illustration of how the common design structure and substitutable feature discovery would function within a feature-based design system.

between CAD models needs to be precomputed as "offline" activity (that could potentially be further reduced by prior segmentation or clustering of the dataset). Such offline preparation of the library would make it feasible to identify substitutable features during an interactive CAD modelling process.

The integration of the proposed CDS and substitutable feature discovery process within a feature-based design system could also enhance the capabilities of feature-based modelling in areas of industrial-feature ontologies and the generation of formal feature semantics. Both attributes that are still required to support feature interoperability for cross-domain and multi-view engineering [19,5]. In this context the library could potentially act as an independent (i.e. neutral) entity in facilitating exchanges between incompatible systems. Since functional commonality is implicit in the substitutable features and could support emerging design technologies related to functional feature modelling (i.e. the integrate geometrical forms, functions, and behaviour in CAD environments) [7,6,20]. Another useful extension of this work could be in the area of intelligent agents used for feature modelling in computer-aided design. Fougères and Ostrosi [8] proposed agents as elementary geometrical and topological objects that possess a knowledge (i.e. awareness) of the context of their application (i.e. local region) in CAD modelling. The concept proposed in Fougères and Ostrosi's work could be further extended using the CDS and substitutable features, which are derived from analysing entire industrial databases. Particularly, the CDS and substitutable features could be developed as agents with associative properties that helps in fusing, expanding, and dividing agents. Such expanded agents could be underpinning a new generation of product development systems.

7. Conclusion and future work

This paper has demonstrated the potential of frequent itemset mining for design applications. Although promising the results presented are based on the analysis of a single dataset of valve designs. Consequently, an investigation of large datasets from other engineering domains is required. The presented work could be expanded to cover other feature types together (such as boss, slot, notch, etc.). Understanding the evolution of the extracted design structures could be studied in-detail if the development histories were available for every component. This progressive development would enable the study of scalability and modification occurring in the design structures.

Future work will focus on assisting engineers in transferring the identified CDS in new design development effectively. The work involves an interactive user interface for CAD design integrated with a knowledge base system that enables predictive design structure and substitutable feature suggestions by actively ensuring compatibility between new design and design structure, and this facilitates reuse as an inbuilt activity in any new product development. Although the proposed approach for the common design structure discovery is described here as being applied to components, the approach could be expanded to cover assembly models. However, to do that additional conditions may need to be added to the five conditions used to identify substitutable features to ensure that compatibility between modified assembly models is assured.

CRediT authorship contribution statement

Gokula Vasantha: Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing - original draft. **David Purves:** Data curation, Formal analysis, Investigation, Methodology, Software. **John Quigley:** Investigation, Funding acquisition, Methodology, Project administration, Resources, Supervision, Writing - original draft. **Jonathan Corney:** Funding acquisition, Methodology, Project administration, Resources, Supervision. **Andrew Sherlock:** Data curation, Formal analysis, Investigation, Methodology, Software. **Geevin Randika:** Data curation, Software, Writing - original draft.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

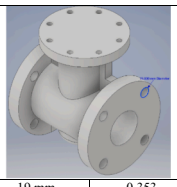
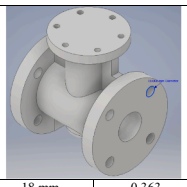
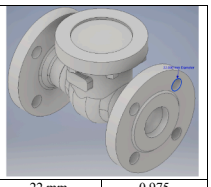
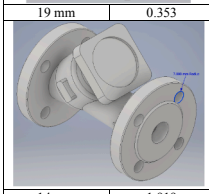
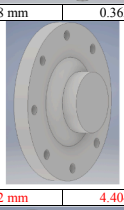
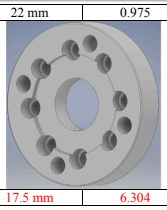
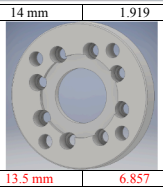
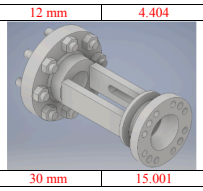
Acknowledgements

This work was supported by the Engineering and Physical Sciences Research Council, UK [grant number EP/R004226/1]. The dataset of hole features extracted from the valves models is available at https://doi.org/10.15129/310393b8-93e1-46ad-b831-74344e030baa.

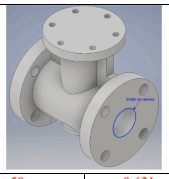
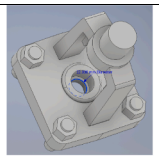
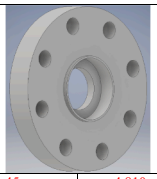
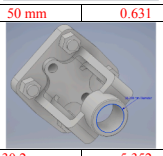
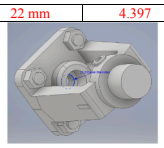
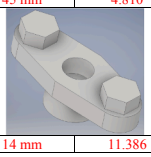
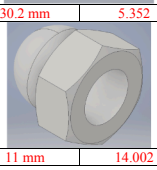
Appendix A

The substitutable hole diameters obtained for 19.05- and 32-mm hole diameters are listed with the K.L. measure score and the associated CAD image.

The best K.L. measure score and the CAD image of a component that contains the substitutable hole diameter feature for 19.05 mm hole diameter (hole diameters that are identified as not suitable for substituting 19.05 mm diameter based on higher K.L. score are highlighted in red colour). (For interpretation of the references to colour in this table legend, the reader is referred to the web version of this article.)

	19 mm	0.353		18 mm	0.363		22 mm	0.975
	14 mm	1.919		12 mm	4.404		17.5 mm	6.304
	13.5 mm	6.857		30 mm	15.001			

The best K.L. measure score and the CAD image of a component that contain the substitutable hole diameter feature for 32 mm hole diameter (hole diameters that are identified as not suitable for substituting 32 mm diameter based on higher K.L. score are highlighted in red colour). (For interpretation of the references to colour in this table legend, the reader is referred to the web version of this article.)

	50 mm	0.631		22 mm	4.397		45 mm	4.810
	30.2 mm	5.352		12 mm	6.585		14 mm	11.386
	11 mm	14.002						

Although the K.L. score for the 50 mm bore diameter is less than value four, it is still not a substitutable feature for 32 mm stem diameter.

References

[1] R. Agrawal, R. Srikant, September. Fast algorithms for mining association rules. In Proc. 20th int. conf. very large data bases, VLDB, 1994, Vol. 1215, pp. 487-499.

[2] A. Baddeley, E. Rubak, R. Turner, Spatial Point Patterns: Methodology and Applications with R, CRC Press, 2015.

[3] J. Bai, S. Gao, W. Tang, Y. Liu, S. Guo, Design reuse oriented partial retrieval of CAD models, Comput. Aided Des. 42 (12) (2010) 1069–1084.

[4] J. Bai, H. Luo, F. Qin, Design pattern modeling and extraction for CAD models, Adv. Eng. Softw. 93 (2016) 30–43.

[5] H. Besharati-Foumani, M. Lohtander, J. Varis, Fundamentals and new achievements in feature-based modeling, a review, Procedia Manuf. 51 (2020) 998–1004.

[6] Z. Cheng, Y. Ma, A functional feature modeling method, Adv. Eng. Inf. 33 (2017) 1–15.
[7] Z. Cheng, Y. Ma, Explicit function-based design modelling methodology with features, J. Eng. Des. 28 (3) (2017) 205–231.
[8] A.J. Fougères, E. Ostrosi, Intelligent agents for feature modelling in computer aided design, J. Computat. Des. Eng. 5(1) (2018) 19–40.
[9] P. Fournier-Viger, A. Gomariz, M. Campos, R. Thomas, Fast vertical mining of sequential patterns using co-occurrence information. In Pacific-Asia Conference on Knowledge Discovery and Data Mining (pp. 40-52). Springer, Cham, 2014.
[10] P. Fournier-Viger, C.W. Lin, A. Gomariz, T. Gueniche, A. Soltani, Z. Deng, H.T. Lam, The SPMF Open-Source Data Mining Library Version 2, in: Proc. 19th European Conference on Principles of Data Mining and Knowledge Discovery (PKDD 2016) Part III, Springer LNCS 9853, 2016, pp. 36-40.
[11] F. Fumarola, P.F. Lanotte, M. Ceci, D. Malerba, CloFAST: closed sequential pattern mining using sparse and vertical id-lists, Knowl. Inf. Syst. 48 (2) (2016) 429–463.
[12] F. Giannini, K. Lupinetti, M. Monti, Identification of similar and complementary subparts in B-rep mechanical models, J. Comput. Inform. Sci. Eng. 17(4) (2017).
[13] R. Huang, S. Zhang, X. Bai, C. Xu, B. Huang, An effective subpart retrieval approach of 3D CAD models for manufacturing process reuse, Comput. Industry, 67 (2015) 38–53.
[14] C. Jackson, M. Buxton, The design reuse benchmark report: Seizing the opportunity to shorten product development, Aberdeen Group, Boston, 2007.
[15] S. Kullback, R.A. Leibler, On information and sufficiency, Ann. Math. Statist. 22 (1) (1951) 79–86.
[16] Z. Li, X. Zhou, W. Liu, A geometric reasoning approach to hierarchical representation for B-rep model retrieval, Comput. Aided Des. 62 (2015) 190–202.
[17] L. Ma, Z. Huang, Y. Wang, Automatic discovery of common design structures in CAD models, Comput. Graph. 34 (5) (2010) 545–555.
[18] D. Paterson, J. Corney, Feature based search of 3D databases, in: ASME 2016 international design engineering technical conferences and computers and information in engineering conference, Am. Soc. Mech. Eng. Digital Collection (2016).
[19] E.M. Sanfilippo, S. Borgo, What are features? An ontology-based review of the literature, Comput. Aided Des. 80 (2016) 9–18.
[20] C. Sen, Feature-based computer modeling and reasoning on mechanical functions, in: International Design Engineering Technical Conferences and Computers and Information in Engineering Conference (Vol. 50084, p. V01BT02A008). American Society of Mechanical Engineers, 2016.
[21] V.B. Sunil, R. Agarwal, S.S. Pande, An approach to recognize interacting features from B-Rep CAD models of prismatic machined parts using a hybrid (graph and rule based) technique, Comput. Indu 61(7) (2010) 686–701.
[22] S. Tao, Z. Huang, L. Ma, S. Guo, S. Wang, Y. Xie, Partial retrieval of CAD models based on local surface region decomposition, Comput. Aided Des. 45 (11) (2013) 1239–1252.
[23] P. Wang, J. Zhang, Y. Li, J. Yu, Reuse-oriented common structure discovery in assembly models, J Mech Sci Technol 31 (1) (2017) 297–307.
[24] X. Yan, J. Han, gspan: Graph-based substructure pattern mining, in: 2002 IEEE International Conference on Data Mining, 2002. Proceedings. (pp. 721-724). IEEE, 2002.
[25] J. Zhang, Z. Xu, Y. Li, S. Jiang, N. Wei, Generic face adjacency graph for automatic common design structure discovery in assembly models, Comput. Aided Des. 45 (8-9) (2013) 1138–1151.
[26] J. Zhang, M.i. Zuo, P. Wang, J.-F. Yu, Y. Li, A method for common design structure discovery in assembly models using information from multiple sources, AA 36 (3) (2016) 274–294.