Research in Engineering Design
https://doi.org/10.1007/s00163-024-00434-w

ORIGINAL PAPER

Evaluating the effectiveness of functional decomposition in early-stage design: development and application of problem space exploration metrics

Jinjuan She[1] · **Elise Belanger**[1] · **Caroline Bartels**[1]

Received: 13 February 2022 / Revised: 3 March 2024 / Accepted: 5 March 2024

Abstract
This paper aims to explore metrics for evaluating the effectiveness of functional decomposition methods regarding problem space exploration at the early design stage. Functional decomposition involves breaking down the main purpose of a complex problem or system into a set of more manageable sub-functions, leading to a clearer understanding of the problem space and its various aspects. While various metrics have been used to evaluate functional decomposition outcomes, little literature has focused on assessing its effectiveness in problem space exploration. To address the gap, this research introduces three metrics for problem space evaluation defined by functional models: quantity of unique functions ($M1$), breadth and depth of the hierarchical structure ($M2$), and relative semantic coverage ratio of the problem space ($M3$). An example study is conducted to illustrate the evaluation process, comparing functional analysis with and without explicit human-centric considerations using a power screwdriver as a case product. The analysis in the example study reveals that the breadth of the hierarchical structure (part of $M2$) is marginally larger in the condition with explicit human-centric considerations (Condition A) compared to the condition without such considerations (Condition B). However, no significant differences are observed in terms of other metrics. The qualitative analysis based on semantic comparisons suggests that Condition A facilitates participants in generating a diverse set of functions supporting user safety requirements more effectively than Condition B. Overall, the example study demonstrates the evaluation process for each metric and discusses their nuances and limitations. By proposing these metrics, this research contributes to benchmarking and evaluating the effectiveness of different methods in promoting functional analysis in engineering design. The metrics provide valuable insights into problem space exploration, offering designers a better understanding of the efficacy of their functional decomposition methods in early design stages. This, in turn, fosters more informed decision-making and contributes to the advancement of functional analysis methodologies in engineering design practices.

Keywords Functional decomposition · Function model · Metric · Problem space exploration

1 Functional decomposition and problem space exploration

Functional decomposition in a design task, also known as functional analysis, is the process or activity that identifies functions of a product to be designed that meet specified customer requirements (Hirtz et al. 2002; Pahl et al. 2007). Representations of functional modeling outcomes vary, ranging from a simple hierarchical function tree (Booth et al. 2015c) to a black box functional structure converting inputs to outputs (Malmqvist 1995; Robotham 2010; Shankar et al. 2020; Patel et al. 2020). This process allows designers to view the design problem based on its primary purpose, breaking it down into manageable sub-functions, and, if needed, their

✉ Jinjuan She
jshe@miamioh.edu

Elise Belanger
belanger@miamioh.edu

Caroline Bartels
bartelc2@miamioh.edu

[1] Department of Mechanical and Manufacturing Engineering, Miami University, Oxford, OH 45056, USA

Published online: 04 April 2024

interconnectivity until each sub-function becomes amenable to simple designs (Ullman 2017). The thorough consideration of functions is pivotal for achieving high-quality design outcomes (Hubka and Eder 2001) and potential redesign of product variants (Wong and Wynn 2023). In addition, an empirical study has shown that function trees help reduce fixation and increase quality during idea generation (Atilola et al. 2016).

However, it has been observed that novice designers often encounter challenges in their functional analysis, displaying limited breadth (focusing on partial design problems) or limited depth (superficial decomposition or lacking insights) (Björklund 2013). To foster problem space exploration, researchers have extensively investigated interventions (Schön 2017; Henderson et al. 2019; Ignacio 2022), with functional analysis emerging as a valuable approach facilitating the understanding of problem space (Gray et al. 2015). Several functional analysis methods have been developed, and some have been integrated into engineering design curriculum (Nagel and Bohm 2011; Booth et al. 2015c; Eisenbart et al. 2017; Yildirim and Campean 2020; Van Eck and Weber 2021; She et al. 2022; Reeling and She 2023). Moreover, recent research has demonstrated the integration of functional analysis with requirements analysis better supports systematic thinking across diverse disciplines, particularly for complex products (Krüger et al. 2023).

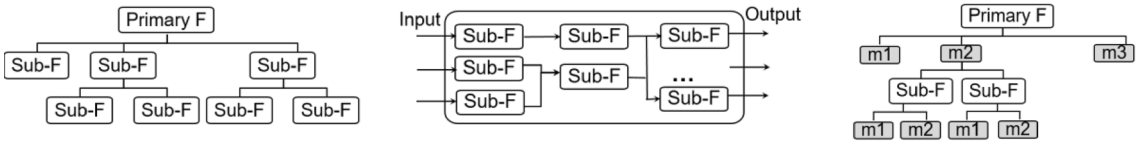
Problem space in early-stage design refers to a landscape in which a problem can be explored and solved, typically encompasses the initial conditions, the desired outcomes, and the possible actions in between (Simon 1973; Goel and Pirolli 1989, 1992). Problem space exploration through function analysis introduces a measure of a designer's holistic thinking in addressing customer needs in a design task, which provides a way to record and communicate primary goals and possible actions needed from the desired products before jumping into the design of physical principles, working principles, embodiment, or details. It requires both abstract (breadth) and concrete (depth) thinking at a functional level, not a morphological level, i.e., form-independent (Pahl et al. 2007; Ulrich et al. 2020). Abstract thinking is core to defining the problem space while concrete thinking of how to support an abstract function contributes to a better conceptual understanding of the function as a whole, i.e., the skeleton elements needed to define a design. This duality in thinking is pivotal in scaffolding subsequent morphological design for elemental functions and synthesis for the whole solution concepts, as illustrated in the descriptive design process model in Majumder et al. (2023). Consequently, this work argues that problem space explored by functional decomposition should be considered when evaluating the effectiveness of a functional analysis method.

However, there remains a notable gap in research, as only very few studies explore metrics to assess the effectiveness of functional analysis methods in problem space exploration in one aspect or another. Previous attempts to measure the shape or geometry of hierarchical function structures in product dissection tasks (Eckert et al. 2011; Booth et al. 2015c) have limitations. The geometry-based approach relies solely on external observations and can be influenced by different interpretations of functions, leading to variations in structural representations. As a result, it fails to capture the full essence of problem space exploration and the effectiveness of functional decomposition methods solely by itself. In the context of product dissection, Eckert et al. (2011) proposed to compare participants' functional representations to a detailed representation for the purpose of pinpointing challenges in identifying functions. Sen et al. (2010a) proposed a general metric to evaluation of information content using an information theoretic approach. While it can be applied to compare competing function models from a syntactic point of view, they also pointed out that a complete evaluation of the semantic content of the model needs to be developed to assess the quality of a function model. Gerick and Eisenbart (2017) undertook a comparison between two functional modeling approaches based on a set of reasoning characteristics they identified. Their examination highlighted the advantages of each approach and suggested that integrating various information led to more comprehensive support for functional analysis. Nevertheless, it is crucial to acknowledge that this recommendation was drawn from the authors' qualitative comparisons in an example decomposition task, without any quantitative measures.

Inspired by the above work, *the authors aim to fill the gap in terms of evaluating problem exploration effectiveness of functional decomposition methods in design tasks. The approach is to explore the evaluation quantitatively, semantically, and qualitatively in the context of functional analysis that supports product design to meet customer requirements.* Specifically, the authors will examine the possibilities of deriving quantitative and semantic metrics of problem space exploration, as well as some suggestions to analyze the breadth and depth of the problem space exploration qualitatively based on the measure. While this paper does not focus on empirically validating the metrics, it uses an example study in order to walk through the exploration and discussion.

It is worth to mention the scopes of this manuscript. First, the evaluation explored does not mean to be comprehensive for all kinds of functional decomposition models. It mainly focuses on models that are represented as a hierarchical structure (see Fig. 1 for some examples), and only considers functions, not the means how to realize functions.

Second, it should be acknowledged that directly assessing problem space exploration is a nuanced endeavor, as there is no single ideal function model even for the same set of customer requirements, and the classification of some



(i) Hierarchical function tree. Showing functions and their sub-functions to address a design problem.

(ii) Function structure graph. Modeling a black box that converts input into output through functions and their connections to address a design problem.

(iii) Function-means tree. Modeling functions and possible means for each function to address a design problem.

Fig. 1 Schematic illustrations of some example functional modeling representations. "F"—functions, "m"—means to realize a function

functions might be ambiguous in a model. Thus, the authors are exploring initial work toward this endeavor through the identification of the characteristics of *relative problem space coverage* that design researchers can use when comparing the effectiveness of different functional modeling methods, or evaluating relative functional thinking abilities within a candidate pool. It can also be used by educators as a tool to curve the assessment of engineering student learning outcomes on this topic.

Next, Sect. 2 delves into the existing literature on the evaluation of functional decomposition, setting the stage for our contribution. In Sect. 3, we introduce our proposed metrics for evaluating the effectiveness of problem exploration in functional decomposition. Section 4 presents an example study, guiding readers through the application, calculation, and analysis of these metrics in an early-stage design context. We then engage in a critical discussion in Sect. 5, where we dissect the subtleties encountered during the evaluation process, acknowledge the limitations of our study, and outline avenues for future research. Section 6 summarizes the main contribution of our work and underscores its practical implications, aiming to inform and advance the practice of functional decomposition in design.

2 Functional decomposition evaluation

As there are infinitely possible functional model variations for any product, evaluating them becomes difficult. While the authors tried their best to review metrics in the literature, they must acknowledge that what is summarized in the paper might not be a complete list. Moreover, not many studies directly compare and quantify the effectiveness of functional analysis methods (Booth et al. 2015c). These metrics are sorted into several categories that evaluate different aspects of functional models. For example, the shape metrics can evaluate level of detail, breadth, and depth, while the rubric measures how well participants adhered to a specific functional analysis method (Booth et al. 2013; Nagel et al. 2015). As this study focuses solely on the functional analysis effectiveness on its immediate outcome (i.e., problem space exploration), metrics that compare the model effect on ideas generated and final designs (e.g., Booth et al. 2015a; Atilola et al. 2016), existing products (McAdams et al. 1999; Dong 2017), or engineer behaviors in the modeling processes (Summers et al. 2017; Patel et al. 2020) are not considered.

2.1 Raw count

Total count is the total number of expressions in a function model (Eckert et al. 2011; Booth et al. 2015c). It does not describe shape, quality, nor account for redundancy or incorrect functions. The raw count alone gives no indication of abstraction level or solution-dependency. In protocol analysis, sometimes the total count is split into written and spoken functions giving more insight into participants' analysis process (Eckert et al. 2011).

2.2 Uniqueness

This category contains both a raw count of unique functions and a ratio with the total number of functions, which describes how efficiently the participant conveyed information and how well they understood the product (Booth et al. 2015c). Uniqueness was determined by semantic similarity rather than exactly identical phrasing to allow for similar functions with different parts and purposes (Booth et al. 2015c). This metric is more subjective, as the semantic similarity relies on evaluators' interpretations about participants' meaning. Similar to the raw count, these uniqueness metrics do not acknowledge if the functions are correct in phrasing or sense.

2.3 Errors

The syntax error metric is a raw count of errors that do not follow the standard "verb-noun" format, whereas the error ratio measures the percentage of errors with respect to the total number of functions generated in a function model. Generic and non-descriptive verbs, such as "to be", are

considered as errors (Booth et al. 2015c). Other errors such as not-solution-neutral (She et al. 2022) or unrelated functions also happen in the modeling process, but are less commonly evaluated. In addition, these metrics only consider the phrasing of the function compared to the rubric that also examines whether participants correctly used the method and that the functions make sense (Tomko et al. 2017).

2.4 Shape

The shape metrics evaluate both the breadth and depth of a hierarchical function model. When interpreting each metric individually, the hierarchical levels relate the levels of abstraction, i.e., how deep the function model goes, while the number of functions on each layer communicates the level of breadth at each abstraction level (Eckert et al. 2011; She et al. 2022) Similarly, in Booth et al. (2015c), the maximum and average geodesic distances were used to quantify the "bushiness" and density of the tree while remaining useful for participants who represented the models with network maps instead of hierarchical tree structures.

2.5 Other measures

This category collects other metrics found in the literature. The first metric is a loose evaluation of overall quality. The categorical evaluation of low, medium, and high quality is simple when dealing with a broad range of models. Still, it allows considerable subjectivity among evaluators by not refining the criteria for low, medium, and high (Booth et al. 2015b). For the energy-flow method, there is a rubric with questions about syntax, nonsensical flows, conservation, and other energy-flow rules. The rubric addresses some aspects of the previous metric categories (such as verb + noun syntax) and evaluates whether the user correctly follows the rules of the method and the quality of the model (Nagel et al. 2015; Tomko et al. 2017). The completeness assessment presents several challenges. The goal is to measure how well the function diagram represents the whole product and all its operations. Eckert et al. (2011) compares participant models to a detailed model, but Booth et al. argues against that method as there is no perfect individual function model to use as a reference (Booth et al. 2013). Furthermore, Kroll uses the fact that so many different models can be created in such a formal analysis process to criticize functional decomposition in general (Kroll 2013). This raises concerns about how to ensure the reference model is adequate and whether or not it is appropriate to use a reference to measure problem exploration. In addition, Gerick and Eisenbart (2017) qualitatively discussed the models' comprehensiveness based on their observations. Hubka and Eder (2001) suggested eight categories of functions based on their purposes (e.g., system purpose functions, assisting functions, production functions, esteem functions). They commented that they can act as checklists to verify if the considerations during designing have been as complete as possible, with the hope of leading to "right-first-time" designing. A metric for computing the information content was also proposed from the syntactic point of view (Sen et al. 2010a). Table 1 summarizes the metrics briefed above and the references that used the metrics.

3 Method: proposed problem exploration effectiveness metrics

In functional modeling research, it is necessary to develop appropriate metrics in terms of which the effectiveness of the modeling methods in enabling problem space exploration can be evaluated and compared. This purpose involves two issues: what to measure and how to measure. In addition, it is reasonable to consider evaluating the processes and the outcome, i.e., the functional model generated. Cognitive processes need to be observed to evaluate the modeling process while a group of designers or individual designers construct functional models using a specific method. It is common to use protocol studies for this purpose (Shah et al. 2000; Patel et al. 2017, 2020; Summers et al. 2017). However, it might be difficult and time-consuming to observe and analyze these cognitive processes (Shah et al. 2000, 2001), which involves a high degree of subjectivity (Dorst and Cross 1995). Therefore, this initial exploration focuses on evaluating the outcome only.

When evaluating problem space exploration, examining how well a function model expands the problem space (breadth) and how well it explores this space (depth) is typical. To prioritize the direct comparison between different function model approaches in problem space exploration, the metrics should be quantitative, objective, and independent of specific vocabularies (i.e., not considering vocabulary requirements such as syntax and solution-neutral, even though they are also important). Considering the nuanced nature of problem space exploration (as discussed in Sect. 1), it is hard to have absolute measures of how well a space is explored. Instead, relative metrics could be useful in comparing functional modeling approaches.

When transitioning from an intangible problem space to a concrete functional representation, assessing the representation using multiple metrics is important. These metrics could include a simple count of functions, a 2D measure of the space represented (similar to length and width in a space), and the specificity of elements within this space (how accurately the generated functions target the desired functions). The number of functions could be an overall indicator, especially the number of unique functions (Metric 1, or $M1$), since it is possible that generating a larger number

Table 1. Metrics for functional model evaluation and references using each metric

Category	Metric	Description	References using metrics*							
			E	B	S	G	N	T	H	Se
Raw count	Total functions	Total number of expressions written on a function model. Repeated or incorrect functions are also counted.	x	x						
Uniqueness	Unique functions	Total number of non-repeated expressions. Rule out repeated expressions.		x	x					
	Efficiency	The proportion of unique functions to total functions.		x						
Errors	Syntax errors	Incomplete functions or functions that do not follow the standard "verb-noun" formal.		x	x					
	Solution neutral	Functions that do not indicate specific solutions, i.e., form-independent.			x					
Shape	Hierarchical layers	Total number of vertical levels present on a function model.	x	x	x					
	Functions on layers	Total number of functions present on each layer across all branches.	x	x	x					
	Maximum geodesic distance	"The longest of the shortest paths on the diagram".		x						
	Average geodesic distance	The average length of all the shortest paths between nodes on the diagram.		x						
Other	Quality	The general quality is rated as high, medium, or low based on an initial impression of clarity, detail, and correct syntax and diagramming. The evaluation depends on the evaluators' subjective perceptions and covers many different criteria at once.		x						
	Rubric	Rate a model based on pre-defined rubrics. Only applicable to the energy-flow method.		x			x	x		
	Checklist	Use categories of functions as checklists to verify that the considerations in design have been as complete as possible. Generic qualitative evaluation.							x	
	Completeness	How well a model explores the target problem. Only qualitative discussions, no actual quantitative output.	x			x				
	Information content	Assess how much information is contained in a model representation from the syntactic point of view.								x

*References are shortened in the table, specifically, "E" is for Eckert et al. (2011), "B" is for Booth et al. (2015b, c), "S" is for She et al. (2022), "G" is for Gericke and Eisenbart (2017), "N" is for Nagel et al. (2015), "T" is for Tomko et al. (2017), and "H" is for Hubka and Eder (2001), "Se" is for Sen et al. (2010a)

of functions might increase the chances of more aspects or more variations of one aspect being considered than fewer functions. In addition, the spatial geometry can be simplified into a box model to represent a top-down hierarchical structure (Metric 2, or *M*2), where "depth" represents the distribution (mode, mean, median, or maximum) of levels, and "breadth" represents the distribution of functions across the levels. Both the count of unique functions and the box model serve as rough estimates of the extent of exploration, with little to no consideration of the semantic meanings of individual functions. To address this, a third metric (*M*3) evaluates the semantic coverage of the exploration: a benchmark function model is established, against which each participant model is compared to calculate a semantic coverage ratio, essentially quantifying how well the problem space is semantically represented. Figure 2 illustrates the interconnections among the three metrics, which we will examine in more detail later. Examples illustrating these metrics are provided in Sect. 4.2 for *M*1 and *M*2 and Sect. 4.3 for *M*3.

3.1 *M*1. Quantity of unique functions

The quantity of unique functions is the total number of unique functions generated, no matter if it is relevant to the design problem or uses the right syntax. A function is considered unique if it is different from others semantically. This will be based on some subjective interpretations; therefore, discrepancies might happen in the counting. Discussions to resolve the discrepancies are expected.

3.2 *M*2. Geometric depth and breadth

Depth is measured by the number of levels. Breadth is measured by the number of functions on a level. They are both discrete quantitative data. If a researcher cares more about the central tendency of a function model, then the summary statistics that represent the typical values can be used: mean, median, or mode (Manikandan 2011a, b) (i.e., depth—mean, median, or mode of the number of functions on a branch in a function tree; breadth—mean, median, or mode of the number of functions on a level in a function tree). If a researcher is interested in the extremes a model can go, then the maximum number of levels or functions on a tree can be considered. However, the statistical maximum is more sensitive and might be biased due to outliers in this context, as it is less likely to analyze and remove outliers for individual function models before measuring the depth and breadth of each function structure. Therefore, measures of central tendency are suggested. Mean is the most well-known average value and is less affected by the number of data available while subject to skewed distribution compared to the median and mode. Median and mode are less accurate than mean when the data size is small, e.g., less than 25 (Hozo et al. 2005). Overall, median and mode are recommended for both Depth and Breadth in *M*2 if a function model is complicated (e.g., has more than 25 levels and more than 25 functions per level) to reduce the effect of possible skewness. Otherwise, mean is recommended to reduce inaccuracy caused by limited data size.

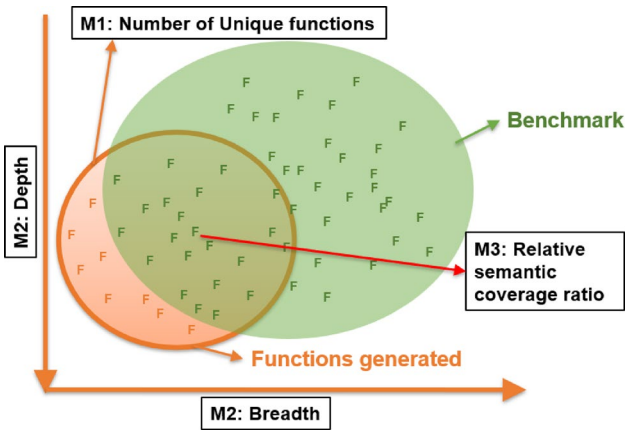


Fig. 2 The relationships between the proposed three quantitative metrics to compare different function model approaches in problem space exploration

3.3 *M*3. Relative semantic coverage ratio

The relative semantic coverage ratio assesses how comprehensively a functional model captures the range of meanings and implications within the problem space relative to a benchmark model. It measures the extent to which the decomposed functions cover the semantic breadth of the user needs and requirements, with the benchmark model as a datum. This ratio would reflect relevance of the functions identified, indicating how well they correspond to the various facets of the problem being solved. Examples are provided in Sect. 4.3 for a better understanding this metric.

First, a benchmark model is needed to evaluate the relative coverage ratio of the problem space semantically. Researchers can collect all functions generated by all participants, affinitize the functions conservatively (put the two into the same group only if their semantic meanings and the abstraction levels are the same), identify unique function categories, and then organize these into a model. Next, experts review the function model against the problem statement and add additional functions to support customer requirements better. Limitations need to be noted. The benchmark function depends on the participant pool's performance and the experts' expertise. Further, there is no single function model even based on the same set of functions, as the order of the functions or the placement of a function on a hierarchical structure might be different. Second, evaluators or raters compare each participant's model against the

benchmark and then count how many benchmark functions appear in a participant's model. Note that one function from a participant's model can only be used once in the comparison and is mapped to the closest benchmark function. Discussions are needed to resolve discrepancies in comparisons. Even though the frequency of a function in a model might make its importance different, such as appearing only once vs. ten times, the frequency is not considered in this evaluation. The main goal of the metric is to evaluate space exploration, i.e., the areas that are considered in exploring the problem. Similar to what other researchers did in externalizing design space considered by teams, only the unique concepts were included in creating the design space map (Gero and Milovanovic 2023). Third, dividing the total count by the total number of benchmark functions leads to a percentage, the quantified semantic relative coverage ratio. The higher the percentage is, the higher the coverage of the space semantically.

Both $M1$ and $M3$ involve a little bit of subjectivity when evaluating if a function is unique ($M1$: 1 = unique, 0 = not unique) or if a function in the benchmark model is included in a participant model ($M3$: 1 = included, 0 = not included). The quality of the metrics is essential to make the comparisons meaningful. In many assessments that involve evaluators or raters, it is typical to use standard inter-rater reliability and inter-rater agreement to measure evaluation quality. The former is about inter-rater consistency in ratings, which quantifies the reliability with a correlation coefficient, such as Cronbach's alpha (Cronbach and Shavelson 2004), with a possible range from 0 to 1.0. A higher coefficient indicates a higher reliability of the evaluations. The latter is about inter-rater consensus, which examines how often the evaluators/raters give the same results, as quantified by the percentages of the same results (Fleenor et al. 1996). When one is interested in the absolute value of the ratings, it is necessary to measure inter-rater agreement (Fleenor et al. 1996). To derive $M1$ and $M3$, the rating values are 0 or 1. Therefore, reporting inter-rater agreements and resolving discrepancies makes more sense. Section 4 will discuss the evaluation process in the context of an example study.

After introducing each measure, the next question is if these measures should be consolidated into an overall effectiveness metric for problem space exploration. The authors would propose not, as there are several problems with such a synthesized metric. First, each of the three is a different type of value on a different scale ($M1$ is a one-dimensional total number, $M2$ can be considered as a two-dimensional pair that describes the length and height of a rectangular shape, and $M3$ is a percentage) and adding them directly makes it difficult to interpret. Normalizing them before adding other ways of synthesizing will make it difficult to understand the meaning of the unified metric. In addition, it is not always the case that all three are applicable to all functional model representations. For example, $M2$ does not make sense for an energy-flow diagram model, while it suits a hierarchical model like a function tree well.

4 Example study walkthrough

4.1 Data from the case study

The authors will walk through the evaluation with the data collected in an experiment reported in She et al. (2022) for a different research purpose. Using functional models in the industry for real-life designs to increase face validity is ideal. However, the extent to which functional modeling is being used in industry may be limited and varied, and access to such information might also be limited for privacy reasons. Therefore, the example study used a made-up design problem solved by groups of novice engineers in a design classroom setting, like how the evaluation metrics for ideation methods were discussed in the design literature (Shah et al. 2000).

This manuscript primarily demonstrates how the metrics can be applied to evaluate functional models created by engineers. This evaluation aims to highlight differences in exploring problem spaces under various conditions. However, the focus of this work is not on any specific conditions, nor does it claim that as its main contribution. The example study details for illustration purposes will be briefly summarized below, but for a comprehensive understanding, readers are encouraged to consult the detailed study in She et al. (2022). This referenced study explores a straightforward approach for incorporating user considerations into functional analysis, addressing challenges often faced by novice engineers at early design stages.

In this example study, 38 sophomore engineering students from a U.S. public university participated. All participants had prior knowledge of functional decomposition. After excluding data from participants who did not provide consent forms, data from 32 respondents (27 male, 5 female) were analyzed. Participants were randomly divided into two groups: one focusing on human-centric condition (Condition A) and a baseline group without this focus (Condition B). Both groups received the same basic information, including a persona and product requirements. However, Condition A also included these requirements within a user workflow, emphasizing user-centric considerations more prominently. As emphasized in She et al. (2022), the requirements provided in the two conditions were exactly the same, except for the representation format, embedded into user workflow (Condition A) vs. a list (Condition B).

The task involved designing a power screwdriver, focusing on decomposing its primary function of "driving screws" into sub-functions (see Fig. 3 for more information). This

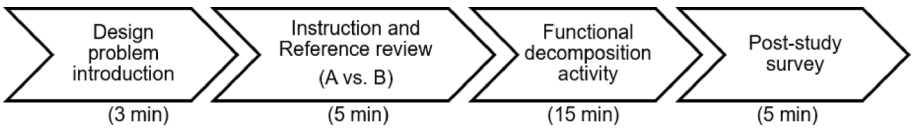
Fig. 3 The design problem associated with the functional decomposition task

Scenario: Imagine that you are going to design a power screwdriver for homeowners who like indoor DIY projects, such as replacing wall covers, installing shelves. Before starting generating design ideas, it is important to understand given customer requirements, and decompose the main function into sub-functions to help more systematic concept generations later on.

Requirements:

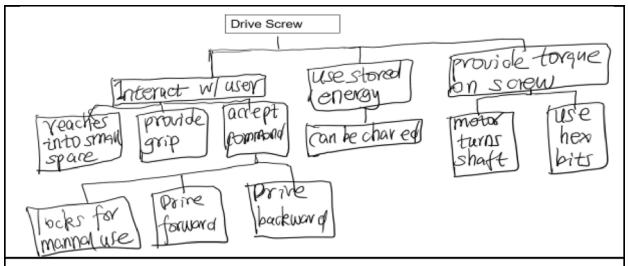
1. Safe to use
2. Can be driven by power and manually
3. Compact for easy storage and transportation
4. Easily rechargeable for long lasting use
5. Cordless
6. Maximum speed limit of 200 rpm
7. Hex drive supports various bits
8. Good for use in normal and tight space
9. Easy grip
10. Able to drive a bit forward and backwards

Fig. 4 The design problem associated with the functional decomposition task



decomposition was to be as thorough as possible, resulting in a function tree where the simple design components could fulfill the lowest level sub-functions. Participants could use any method or combination of methods learned in class (such as energy flow, top-down, and enumeration) for generating these function tree structures. Importantly, each sub-function was to be represented by "verb-noun" pairs and be independent of specific forms. A post-study survey gathered additional data on each participant's background, including their familiarity with functional analysis, challenges faced during the process, and demographic information. Figure 4 depicts an overview of the study process. For a detailed description of the stimuli used and a comprehensive overview of the study process, readers should refer to She et al. (2022) and the figures included in our manuscript.

All function trees were transcribed into spreadsheets. One researcher reviewed the transcribed data and the raw data to ensure no information loss or change. See Fig. 5 for an example of the raw functional model data and the transcribed data.



(i) An example function model raw data

Func ID	Function Name	Level
B18.1	interact with user	1
B18.1.1	reaches into small space	2
B18.1.2	provides grip	2
B18.1.3	accepts commands	2
B18.1.3.1	locks for manual-use	3
B18.1.3.2	drive forward	3
B18.1.3.3	drive backward	3
B18.2	use stored energy	1
B18.2.1	can be charged	2
B18.3	provides torque on screw	1
B18.3.1	motor turns shaft	2
B18.3.2	uses hex bits	2
Summary	Total number of levels	3
	N of functions on the 1st	3
	N of functions on the 2nd	6
	N of functions on the 3rd	3
	N of functions on the 4th	0
	N of functions on the 5th	0

(ii) Transcribed data

Fig. 5 An example functional decomposition generated by Participant B18 and its corresponding transcribed data

4.2 Quantity evaluations: *M*1 and *M*2

4.2.1 *M*1. Quantity of unique functions

In evaluating unique functions, the research team deemed the coding of 1 to represent unique functions and 0 to represent repeated functions. A function is considered unique if it has a different meaning from the other functions on the function tree. Two researchers separately coded the metric and then compared their coding to discuss and correct the discrepancies. One example case that caused different evaluations is due to the commonly tied function pairs. For example, some participants wrote power on/off as one function, while some wrote them as separate functions, power on and power off. The functions on and off are typically coupled

together; therefore, no matter if it is represented as one function ("on/off") or two functions "on" and "off"), it is only counted as one unique function for equal comparison. Based on the criteria defined, $M1$ for the example function tree in Fig. 6 is 14. The functions "rotate w/ hand" and "rotate w/ power" were only counted once while "rotate forward" and "rotate backward" were considered as one unique function "rotate forward/backward".

4.2.2 *M*2. Depth and breadth of geometry shape

To measure the geometry shape of a functional decomposition tree, statistics of central tendency are used as recommended in Sect. 3. The number of functions on each level and each branch are counts, and can be considered discrete quantitative data. If the tree structure is simple, using the median to calculate depth and breadth is less representative, and the mean is preferred. For example, Fig. 6 illustrates a small-sized function tree, which has 11 branches with the number of functions along each branch as 2, 2, 3, 3, 3, 3, 2, 2, 2, 2, and 2 respectively. Then, the mean depth of the entire tree is $(2 \times 7 + 3 \times 4)/11 = 2.36$. Note that the first function, drive screws, was given (not generated by participants). Therefore, it was not counted. For breadth, the number of functions is counted at each level, with 4, 9, and 4 at each level. Then, the mean of breadth is $(4 + 9 + 4)/3$=5.67.

4.3 Semantic coverage ratio evaluation: *M*3

4.3.1 Generate a benchmark model

In two steps, a benchmark functional decomposition of the power screwdriver was developed as a baseline for relative problem space coverage evaluation. First, individual functions mentioned by all participants were affinitized by the authors based on their semantic meaning. It is noted a limitation of this study, even though the authors were unaware of the conditions of the data they affinitized. Ideally, the affinity activity could be conducted by independent people who understand the case product to reduce researcher bias. Fig. 7

(a) Affinitize the participants generated functions

(b) Group categorized participant functions into a new tree.

Fig. 7 Generate the first version of the benchmark function model based on the participant data

illustrates the affinity activity and the outcome. In the activity, teams recorded all participant functions on individual post-it notes, then grouped them for similarities and gave each group a name. Two functions were grouped if they (or the closest functions they imply) referred to the same function semantically at approximately the same abstract level (e.g., *control speed* vs. *limit RPM to 200*). Two functions were not grouped if they represented different hierarchical levels of a group of functions (e.g., *provide controls* vs. *control power status on/off*) or meant different design spaces for a higher level function (e.g., *control power status on/off* and *provide direction choices forward/backward* were not grouped, even though they were both sub-functions of *provide controls*). Each group was given a function name to represent the group. The name was picked from the participant phrases and refined by the authors to meet solution-neutral criteria (i.e., how a function is expressed does not indicate a solution) and syntax correctness (i.e., functions are described in verb and noun pairs). Next, the authors put

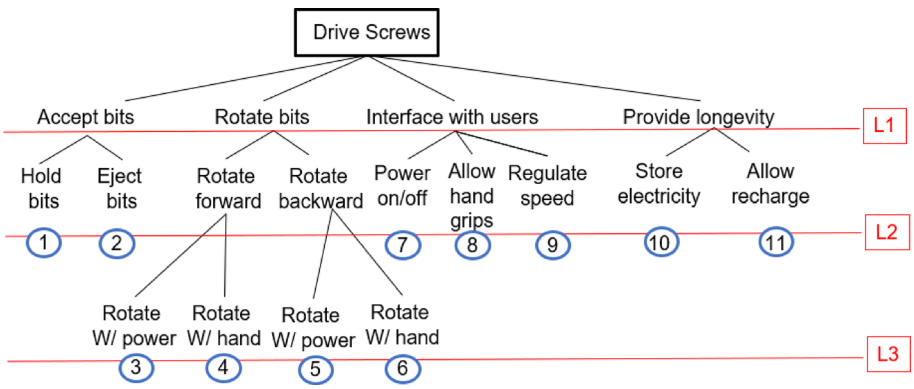


Fig. 6 An example function tree structure with 11 branches (indicated by a blue circle at the bottom of each branch), and three levels (indicated with a red line and a red rectangle on the right of each level). Depth is calculated as the mean number of functions across all 11 branches, and breadth is calculated as the mean number of functions across all three levels

the newly generated function names into a tree diagram, serving as the benchmark function model against which to calculate each participant's functional model's relative semantic coverage ratio.

It is possible that participant data missed some information based on their understanding in a limited time, even with all participant data considered. In the second step, to further solidify the functional model based on the best knowledge of the authors, the research team refined the benchmark tree before evaluating the space exploration coverage. The team recapped the prompts given to the participants and noted down function trees they would each individually generate to the best of their knowledge. They were asked to add additional functions if they identified anything new to the benchmark generated in Step 1. Then, each of the authors in She et al. (2022) reviewed a quarter of the participants' function trees as assigned, and cross checked if any functions were missing in the benchmark tree. Discussions were conducted at the end to consolidate the findings. Note that some functions from participants' data were removed on the benchmark tree purposefully if they were not relevant to the customer requirements or the product under design as it is not fair to penalize others who did not include irrelevant or random functions for coverage completeness measures. For example, "add a magnetic property of the bit" was removed, as it is about bit design, and not relevant to a power screwdriver design or the given customer requirements. Figure 8 depicts the final benchmark function tree.

4.3.2 Evaluate each participant's functional model against the benchmark model

Assessment of the problem space relative coverage compared a participant's tree to the benchmark tree based on functions' semantic meaning. The assessment was documented in a template, as shown in Fig. 9. The left columns tabulated individual functions of the benchmark model. Each participant's function tree was reviewed against the benchmark tree by two evaluators (one evaluator is not the author, but noted in the acknowledgment). A cell was marked as "1" if one function from the participant's model could be mapped to the benchmark function in the corresponding

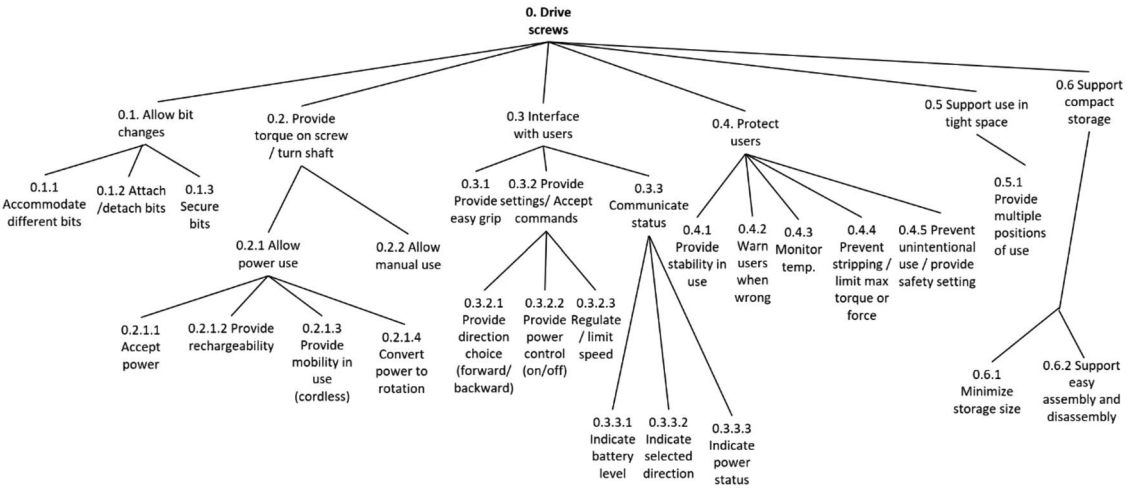


Fig. 8 Benchmark function tree generated by the authors based on all participant data

Fig. 9 A screenshot of the partial completeness evaluation form

Func ID	Individual Functions in Benchmark (See benchmark tree on next tab)	Participant ID					
		B1	B2	B3	B4	B5	....
0.1	**Allow bit changes**	0	0				
0.1.1	Accommodate different bits	1	0				
0.1.2	Attach/detach bits	0	0				
0.1.3	Secure bits	0	0				
0.2	**Provide torque on screw / turn shaft**	1	0				
0.2.1	Allow manual use	0	1				
0.2.2	Allow power use	1	1				
0.2.2.1	Accept power	0	0				
0.2.2.2	Store power	1	0				
...	...	...	...				

row, marked as "0" otherwise. One function could only be used once, and mapped to the benchmark function that had the closest abstraction level. To calibrate the understanding, all evaluators reviewed and discussed two function trees together before formal evaluation. The evaluations were compared to identify differences. Differences were discussed and resolved as a group.

4.3.3 Calculate the relative semantic coverage ratio (*M*3)

Semantic coverage was quantified by a percentage of benchmark functions covered in a participant's model, i.e., Semantic Coverage = total count of functions mapped to benchmark/32 ("32" represents the total number of functions in the benchmark model in this example study). Comparing the function tree in Fig. 6 generated by a participant to the benchmark function tree in Fig. 8 based on the semantic meaning of each function, we found that 11 out of the 32 functions in Fig. 8 appeared in Fig. 6. Therefore, M3 for the function tree in Fig. 6 is 11/32=34%.

4.4 Metric analysis

4.4.1 Statistical tests between the groups

The metrics calculated in Sects. 4.2 and 4.3 can be analyzed and compared to quantitatively test the effectiveness of different functional analysis methods. Before analysis, outlier data of each metric were identified and removed. The selection of statistical tests was meticulously tailored to match the data distribution and the specific assumptions inherent to each test. ANOVA tests were employed for metrics *M*2 Breadth and *M*3 Relative Coverage, as these metrics displayed a normal distribution and exhibited homogeneity of variance. Conversely, for other metrics that demonstrated a non-normal distribution, Wilcoxon Rank Sum tests were utilized to appropriately address these data's distinct characteristics.

It should be noted that conducting multiple tests on identical datasets simultaneously increases the risk of encountering a Type I error. The Bonferroni correction is often recommended as a solution, adjusting the level of statistical significance in direct relation to the quantity of hypotheses tested (for instance, adjusting the significance level to 0.0125 from 0.05 when testing four hypotheses) (Emerson 2020). Nonetheless, this reduction in Type I error consequently heightens the possibility of a Type II error, which involves mistakenly accepting a false hypothesis as true (Bender and Lange 2001). The appropriateness of p value adjustments remains a contentious topic within statistical discussions (Perneger 1998; Bender and Lange 2001; Narum 2023). Although it introduces complexity and ambiguity, we took Narum's suggestions to report raw p values (Narum 2023), followed by a discussion within the text. This strategy aims to provide readers with a thorough comprehension of the data, allowing them to draw their own conclusions, particularly as this study demonstrates the analysis of various metrics. Recognizing that conventions regarding significance levels may vary across different fields is also important. Traditionally, significance levels for hypotheses are categorized as weak (0.1), moderate (0.05), and strong (0.01) without adjustments. However, upon applying corrections for multiple comparisons, these levels are recalibrated by dividing them by the total number of hypotheses tested, ensuring a more rigorous evaluation of statistical significance.

In this example study, the comparative impact of the two functional analysis methods (A vs. B) on the number of unique functions, geometric depth of the function structure, or relative semantic coverage ratio was trivial. However, a potential statistical difference in the breadth of functional decomposition was observed, even though at a marginal level ($F(14, 17) = 3.45$, p value < 0.1), depending on the significance level chosen as discussed above. See Table 2 for a summary.

4.4.2 Qualitative analysis of individual functions and groups of functions

First, the total frequency of a function that appeared in a condition was counted and then normalized according to the number of participants (i.e., frequency of mentioning a function in a condition/total number of participants in the condition). The process was repeated for all functions in the benchmark model. Next, individual and group functions

Table 2 Descriptive statistics summary

Metrics	Condition A		Condition B		Statistics	p value
	Mean ± SD	n	Mean ± SD	n		
*M*1: unique	14.10 ± 5.03	14	12.30 ± 2.93	18	W = 151	0.35
*M*2: breadth: MeanNFperL	5.17 ± 1.10	14	4.53 ± 0.82	17	F = 3.45	0.07
*M*2: depth: MeanOfDepth	2.05 ± 0.15	11	2.24 ± 0.40	17	W = 73	0.19
*M*3: relative coverage	0.31 ± 0.14	14	0.28 ± 0.08	18	F = 0.43	0.51

SD standard deviation, n number of observations used in analysis after removing outliers, *F* ANOVA test, *W* Wilcoxon rank sum test

were analyzed in two steps: (1) overall observations across both conditions and (2) relative comparisons between the conditions. Overall observations help evaluate the challenges in the chosen design problem, while the relative comparisons further differentiate the different performances between the conditions and support the observable effect of the other functional analysis methods on the individual functions or groups of functions. The paragraphs below will demonstrate the findings from the example study in these two steps.

4.5 Overall observations

It is helpful to review the data as a whole to identify which functions were recognized by most of the subjects and which functions were often missed in subjects' function analysis in the context of a design task. Figure 10 depicts the total frequency mean of a function identified per subject in both conditions. The frequency of a function identified could indicate its easy or difficult identification. This helps to evaluate the design problem and customer requirements: identifying obviously included or easy-to-be-missed functions and then mapping these to the needs indicates the challenging requirements to meet in the early design stage. The smaller the sum, the more difficult, or not obvious, to identify the corresponding functions overall across the two conditions, and vice versa. This information will further help to tailor design research to specific types of requirements or groups of needs to support a more comprehensive exploration of the design space to meet customer requirements better. For example, in this case study, functions that are associated with interface with users (e.g., 0.3.3—communicate status/information, 0.3.3.2—indicate selected direction, 0.3.3.3—indicate power status), protect users (e.g., 0.4.3—monitor temperature, 0.4.4—prevent stripping), and support use in tight space (e.g., 0.5.1—provide multiple positions of use) have obvious low occurrence in summary. It indicates that participants might have more difficulties in coming up with functions to support user interaction-related requirements (such as user-friendliness and accessibility) while less so for the device-centric functions (e.g., 0.1.1—accommodate different bits, 0.2.1—allow poser use, 0.2.2—allow manual use). This observed pattern calls for more research in supporting engineers, at least novice engineers, in functional analysis that considers user interactions.

4.6 Relative comparisons between the two conditions

A relative comparison between the two conditions was conducted to understand which condition led to more frequent identification of specific functions, both individually and in groups. Such analysis can reveal patterns about the ease or difficulty of identifying certain functions. Figure 11 in our manuscript offers a qualitative comparison from the example study, where orange bars represent the mean differences and green boxes categorize functions as per the benchmark model shown in Fig. 8.

In this comparison, larger values indicate that a function was identified more frequently on average in Condition A compared to Condition B, and vice versa. It was observed that participants in Condition A more often identified functions such as "0.1.2-attach/detach bits," "0.1.3-secure bits," "0.2.1.1-accept power," "0.2.1.4-convert power to rotation," and "0.3.3.1-indicate battery level." Conversely, participants

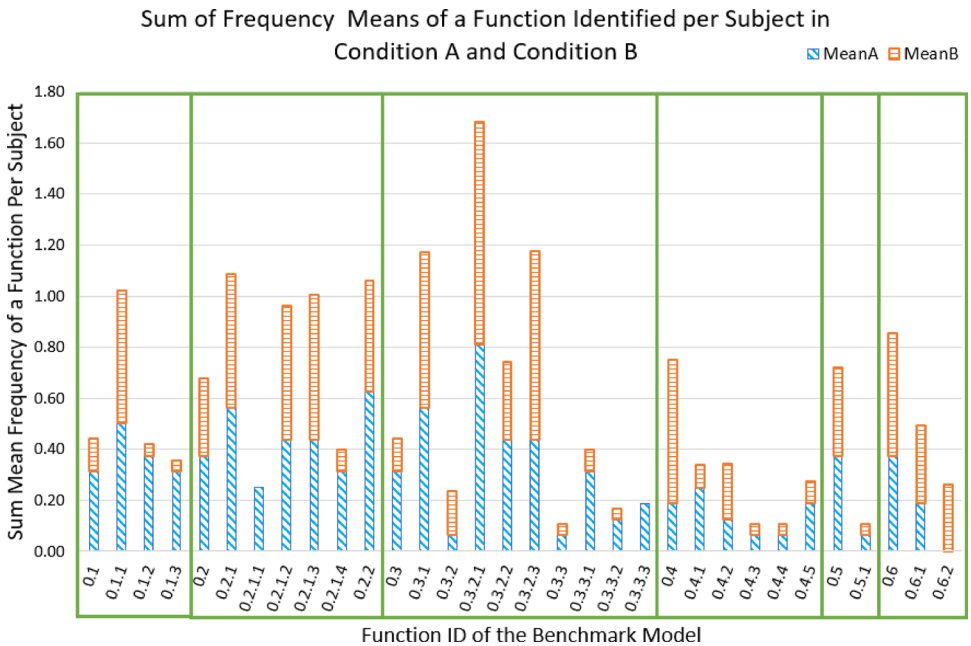


Fig. 10 The sum of mean frequencies of a function identified in Conditions A and B. The smaller the sum, the more challenging it is to identify the corresponding functions overall across the two conditions (e.g., functions 0.3.3, 0.3.3.2, 0.3.3.3, 0.4.3, 0.4.4, and 0.5.1), and vice versa, the larger the sum, the easier to identify the corresponding functions (e.g., functions 0.1.1, 0.2.1, 0.2.2, 0.3.1, 0.3.2.1, and 0.3.2.3)

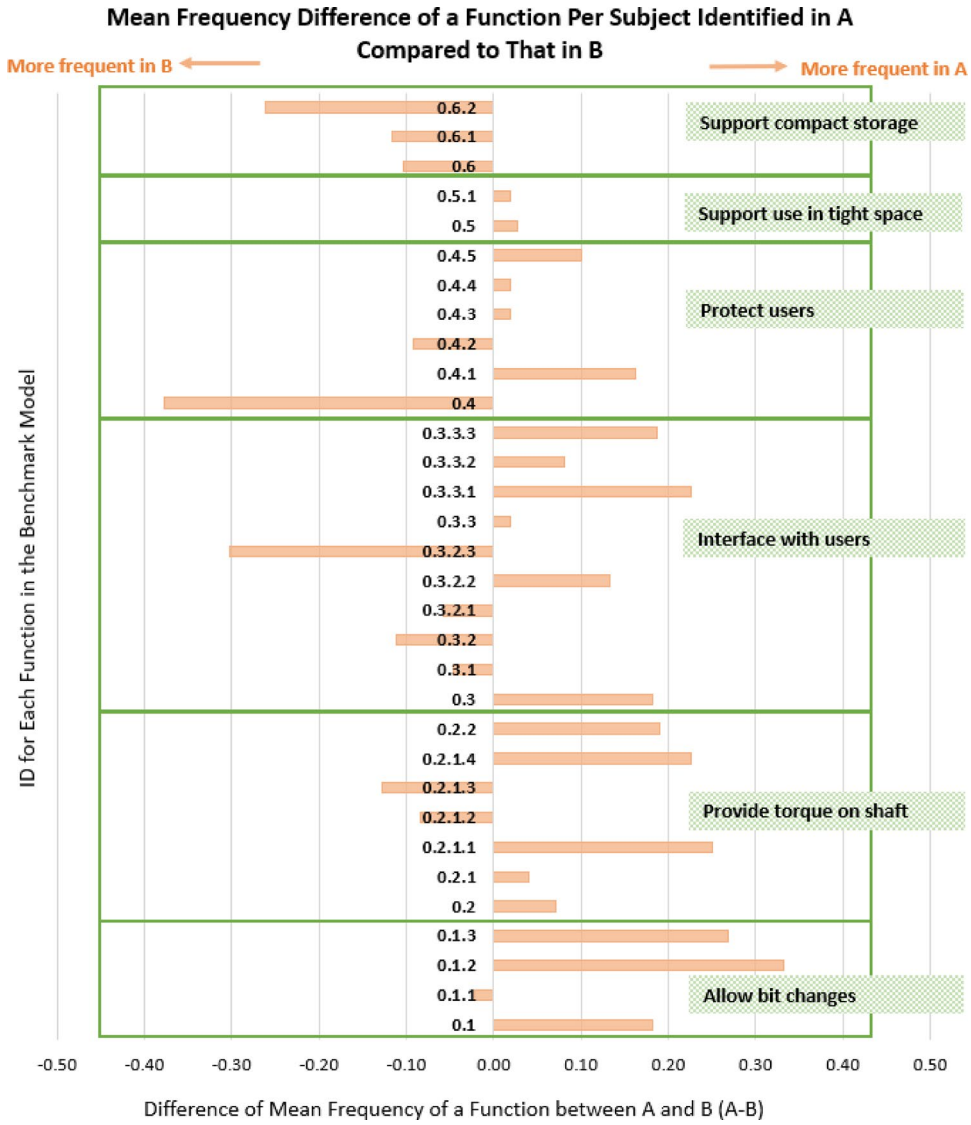


Fig. 11 Mean frequency difference of each individual function identified per subject in Condition A compared to Condition B. Orange bars indicate the mean difference. Green boxes indicate the function categories according to the benchmark model shown in Fig. 8. The larger the value (towards to the right side) means the more frequently that a function is identified on average by the subjects in Condition A than in Condition B (e.g., 0.1.2, 0.1.3, 0.2.1.1, 0.2.1.4, 0.3.3.1), vice versa (e.g., 0.3.2.3, 0.4, 0.6.2). Closer to zero means that the two conditions have almost equal performance on that function (e.g., functions 0.5, 0.5.1, 0.4.4, 0.4.3). See Fig. 8 for the meaning of each function ID

in Condition B were more adept at identifying functions such as "0.3.2.3-regulate or limit speed," "0.4-protect users," and "0.6.2-support easy assembly and disassembly." Functions that scored close to zero indicated almost equal performance between the two conditions, examples being "0.5-support use in tight space," "0.4.4-prevent stripping," and "0.4.3-monitor temperature."

Moreover, the pattern differences also reflected performance variations at different levels of abstraction. For instance, in the category of user protection, participants in Condition B more frequently identified the higher level function "0.4-protect users," but less so its sub-functions. This suggests a more abstract approach to fulfilling the user requirement of safety. In contrast, participants in Condition A identified a wider range of functions supporting user safety, indicating a more detailed and varied approach. This observation implies that different functional analysis methods might favor certain aspects; for example, the method in Condition A seems more effective in helping generate a diverse set of functions that support user safety requirements in functional analysis compared to the method in Condition B.

5 Discussion

This research explores metrics that compare functional models in hierarchical structures. The metrics are helpful for controlled studies that evaluate and especially compare the effectiveness of different functional decomposition methods in early-stage design tasks. Compared to other tasks such as product dissection or reverse engineering, early conceptual design is more driven by customer requirements and needs, with less concrete information about the product. In response to these differences, this paper proposes to measure the problem space exploration quantitatively, semantically,

and qualitatively through three metrics (the number of unique functions, geometric breadth and depth, and relative semantic coverage ratio) and qualitative analysis. The paper used an example study to explain further how to calculate and analyze each.

Uniqueness is a common measure in engineering design as the unique features, forms, or functions differentiate products, affecting customer adoption preferences (Lee et al. 2018). Researchers have used the quantity of uniqueness and the ratio of unique to total functions to measure the richness of the functions generated (Booth et al. 2015c) but did not discuss the results in-depth as they were insignificant in their comparison studies. Similarly, the statistical evaluation of uniqueness in the example study was neither significant between human-centric functional analysis (Condition A) nor baseline functional analysis (Condition B). Considering that uniqueness is a metric that is easy to assess with less effort and is less tied to the specific representation format of the functional models, it is still recommended to include it as one quantitative metric. Uniqueness provides a one-dimensional assessment of the problem space explored. Other metrics are needed for in-depth analysis.

The second metric is the *geometric breadth and depth*. This is more restricted to functional models represented as hierarchical structures, such as function trees. It is suggested that measures of central tendency, such as mean, median, or mode for breadth and depth, be used instead of maximum counts to reduce biases due to outliers; see Sect. 3 for detailed reasoning. Selection of the specific central tendency measures depends on the data distribution, complexity of the models, and effort needed to obtain the metrics. When the function models are not complicated (e.g., fewer than 25 branches or levels), the median or mode of the limited data is not accurate; instead, the mean number of branches and the mean number of levels are recommended. If the models are complicated, the median or mode can be used for efficiency and robustness, and they can be easily calculated either graphically or with the help of programming using a predefined numeric coding structure of functions (e.g., Fig. 8).

In the example study, we chose a simple problem strategically aiming to illustrate the "bare bones" steps and effectiveness of these metrics in a controlled manner. All function trees under evaluation have fewer than 25 branches and levels in the study. Therefore, means were chosen for the depth and breadth measures. The comparison on $M2$ found that Condition A generated function trees that might be broader than Condition B (depending on how strict the significance levels are), while there was no significant difference in the depth of the two conditions (Sect. 4.4.1). This indicates that the method in Condition A has the potential to help expand the breadth of functional decomposition in the design tasks. It might be beneficial in later morphological design stages to enhance the variety of ideas generated and thus expand the design space. The geometric breadth and depth further help space exploration assessment by revealing the outcome from two dimensions, breadth, and depth, which are both crucial factors for the subsequent designs. Note that it might be subject to how modelers put an individual function onto a hierarchical structure, even though it is expected that more abstract functions should be at higher levels and sub-functions should be at lower levels.

This manuscript proposed how to generate a benchmark function model by aggregating participants' input, researchers' expertise, and given requirements and further expanded the qualitative analysis to a quantitative metric–*relative semantic coverage ratio*, which can be used to compare the effectiveness of different functional modeling methods. The ratio supports researchers in analyzing semantic coverage quantitatively. Possible qualitative analyses based on the comparisons were also illustrated in Sect. 4.4.2. For example, in the example study discussed in this paper, participants in both conditions (novice designers) struggle to identify user interaction-related functions. Moreover, Condition B identified protect users more commonly at a higher abstraction level, while Condition A identified more sub-functions to protect users. The observations revealed possible areas in which novice engineers need support (e.g., identifying user interaction-related functions), and also, the functional analysis method in Condition A might have helped participants better decompose a higher level function (protect users) to more sub-functions that are easier to design (e.g., provide stability, monitor temperature, prevent stripping, and prevent unintentional use). It was acknowledged that the lack of details for a requirement or function prevents its proper utilization throughout the design process (Shankar et al. 2020). In future assessments, the scope of problem space coverage may be further refined into distinct coverage ratios. These ratios will align with specific categories of requirements intended for examination in functional analysis. In addition, integrating Design for X (DFX) tools will enable a more focused enhancement of the design, such as design for circularity and durability (Mesa 2023; Mesa and González-Quiroga 2023). Thus, quantitative comparisons on each requirement category are possible if a specific category is of interest to the researchers.

While the proposed metrics provide a multi-dimensional view of problem space exploration, there are some limitations to consider. The benchmark model's subjectivity and potential omission errors may affect the accuracy of the evaluation. In addition, the effort required to derive the relative coverage ratio is high, warranting further research to make this metric more cost-effective, potentially leveraging rubric or checklist approaches or AI algorithms. Furthermore, functions in a hierarchical structure have different importance in shaping a design space. For instance, the absence of a leaf function might

lead to a missing function of the product, whereas missing a mid-branch function still allows for the implementation of leaf functions. For another instance, missing a top-branch function (e.g., allow bit changes in Fig. 8) might lead to missing entirely unexplored aspects of the design. Using equal weights (not accounting for the distinct importance of each function) helps make a fast and frugal assessment, albeit potentially compromising accuracy. Future work should delve into enhanced coverage assessment that incorporates weighted considerations.

Future research should also broaden its scope to encompass a variety of products and functional analysis methods to refine problem space exploration assessment metrics. Delving into the nuances of function density in complex products and integrating the analysis of function repetition and frequency will enrich the assessment's scope. It is equally important to consider the evolving nature of design, adapting to changes in problem definitions and shifting requirements, which adds a critical layer of depth to the evaluation. In addition, precise definition, thorough exploration, and robust validation of effectiveness metrics are important to advancing functional analysis method development and assessment. Extending the relevance of these metrics to different methods and accounting for outlier scenarios will enhance their utility.

In addition, there might be new opportunities for further enhancing the metrics by integrating the geometric and semantic metrics proposed here with the topological, graphical, and syntactic metrics found in existing literature (Sen et al. 2010a, b; Mathieson et al. 2012). For example, the information content metric measures relative information content, and information density based on the topological information, and can be used to evaluate the rate of the evolution of the design as well (Sen et al. 2010a, b). The integration of these work might lead to a more comprehensive and nuanced approach to metric development.

The real-world testing of these metrics through empirical research and industry practice will ground theoretical analyses in practical reality. Comparing the problem exploration strategies of seasoned engineers and novices through these metrics can shed light on enhancing educational methods and developing supportive tools. Understanding the metrics' applicability in different functional analysis tasks will guide tailoring approaches to specific design contexts.

Finally, optimizing the evaluation process through automation or semi-automation will streamline functional analysis evaluation and enhance productivity. Addressing these areas of future research will significantly contribute to advancing functional analysis methodologies and their broader implementation.

6 Conclusion remarks

6.1 Summary

This paper proposes a multi-dimensional approach to evaluate problem space exploration in functional modeling during early-stage design, where the abstract nature of customer requirements and needs makes problem space exploration an important topic for design innovations. We propose three metrics: the number of unique functions, the geometric breadth and depth of the hierarchical structure, and the relative semantic coverage ratio, supplemented by qualitative analysis. These metrics provide a foundational assessment framework of functional decompositions, providing insights that are both broad in scope and specific in application and thus allowing for a deeper understanding of how different decomposition methods affect the exploration of the problem space. Through an example study, the paper demonstrates the evaluation process for each metric and discusses its nuances, limitations, and potential applications. In the example study, the three metrics together assessed the problem space exploration in each functional analysis condition more comprehensively and provided multiple perspectives to examine the differences between the conditions. For example, the functional analysis with explicit human-centric considerations (Condition A) could broaden the problem space considered, even though there was no significant influence on the depth of the space, the total number of unique functions, or the semantic coverage ratio. The qualitative semantic comparison also reveals the strength of Condition A in helping participants generate a diverse set of functions supporting user safety requirements.

6.2 Practical implications

The proposed metrics for assessing the effectiveness of problem space exploration in functional decomposition offer tangible benefits across different facets of engineering design. Using these metrics, design teams can critically assess their functional decomposition methods and choose more effective methods, thereby streamlining project workflows. Instructors can incorporate these metrics (e.g., $M1$ and $M3$) into design courses in education settings, allowing them to quantitatively assess students' functional decomposition. This can help give students specific feedback by providing a clear framework for evaluating the thoroughness of their functional analyses. The metrics can also inspire the development of software tools that aid designers in functional decomposition by providing criteria for software algorithms to assess and

improve problem space exploration. In essence, these metrics provide a quantifiable means to assess the otherwise qualitative process of problem space exploration, thereby enhancing decision-making and supporting the creation of more innovative and successful methods to facilitate problem understanding.

Acknowledgements Mr. Hunter Reeling's contribution in the data evaluation is greatly appreciated.

Funding Funding was provided by Miami University Faculty Startup Fund.

Data availability Not applicable for our work.

References

Atilola O, Tomko M, Linsey JS (2016) The effects of representation on idea generation and design fixation: a study comparing sketches and function trees. Des Stud 42:110–136. https://doi.org/10.1016/j.destud.2015.10.005

Bender R, Lange S (2001) Adjusting for multiple testing—when and how? J Clin Epidemiol 54:343–349. https://doi.org/10.1016/S0895-4356(00)00314-0

Björklund TA (2013) Initial mental representations of design problems: differences between experts and novices. Des Stud 34:135–160. https://doi.org/10.1016/J.DESTUD.2012.08.005

Booth JW, Bhasin AK, Ramani K (2015a) Art meets engineering design: an approach for reducing sketch inhibition in engineers during the design process. In: Proceedings of the ASME design engineering technical conference, Boston, MA, Aug. 2–5. American Society of Mechanical Engineers Digital Collection

Booth JW, Bhasin AK, Reid TN, Ramani K (2015b) Empirical studies of functional decomposition in early design. In: ASME 2015 international design engineering technical conferences & computers and information in engineering conference. Boston, MA, Aug. 2–5

Booth JW, Reid T, Ramani K (2013) Understanding abstraction in design: a comparison of three functional analysis methods for product dissection. In: Proceedings of the ASME design engineering technical conference. Portland, OR, Aug. 4–7

Booth JW, Reid TN, Eckert C, Ramani K (2015c) Comparing functional analysis methods for product dissection tasks. J Mech Des Trans ASME 137:081101. https://doi.org/10.1115/1.4030232

Cronbach LJ, Shavelson RJ (2004) My current thoughts on coefficient alpha and successor procedures. Educ Psychol Meas 64:391–418

Dong A (2017) Functional lock-in and the problem of design transformation. Res Eng Des 28:203–221. https://doi.org/10.1007/s00163-016-0234-3

Dorst K, Cross N (1995) Protocol analysis as a research technique for analysing design activity. In: Proceedings of the ASME design engineering technical conference. American Society of Mechanical Engineers Digital Collection, pp 563–570

Eckert C, Alink T, Ruckpaul A, Albers A (2011) Different notions of function: results from an experiment on the analysis of an existing product. J Eng Des 22:811–837. https://doi.org/10.1080/09544828.2011.603297

Eisenbart B, Gericke K, Blessing LTM, McAloone TC (2017) A DSM-based framework for integrated function modelling: concept, application and evaluation. Res Eng Des 28:25–51. https://doi.org/10.1007/s00163-016-0228-1

Emerson RW (2020) Bonferroni correction and type I error. J vis Impair Blind 114:77–78. https://doi.org/10.1177/0145482X20901378

Fleenor JW, Fleenor JB, Grossnickle WF (1996) Interrater reliability and agreement of performance ratings: a methodological comparison. J Bus Psychol 10:367–380

Gericke K, Eisenbart B (2017) The integrated function modeling framework and its relation to function structures. AI EDAM 31:436–457. https://doi.org/10.1017/S089006041700049X

Gero J, Milovanovic J (2023) The situatedness of design concepts: empirical evidence from design teams in engineering. Proc Des Soc 3:3503–3512. https://doi.org/10.1017/PDS.2023.351

Goel V, Pirolli P (1992) The structure of design problem spaces. Cogn Sci 16:395–429. https://doi.org/10.1207/s15516709cog1603_3

Goel V, Pirolli P (1989) Motivating the notion of generic design within information-processing theory: the design problem space. AI Mag 10:19–19. https://doi.org/10.1609/AIMAG.V10I1.726

Gray CM, Yilmaz S, Daly S et al (2015) Supporting idea generation through functional decomposition: an alternative framing for design heuristics. In: Proceedings of the 20th international conference on engineering design (ICED 15), vol 1: Design for Life, pp 1–10

Henderson D, Jablokow K, Daly S et al (2019) Comparing the effects of design interventions on the quality of design concepts as a reflection of ideation flexibility. J Mech Des Trans ASME. https://doi.org/10.1115/1.4042048/368480

Hirtz J, Stone RB, McAdams DA et al (2002) A functional basis for engineering design: reconciling and evolving previous efforts. Res Eng Des 13:65–82. https://doi.org/10.1007/S00163-001-0008-3

Hozo SP, Djulbegovic B, Hozo I (2005) Estimating the mean and variance from the median, range, and the size of a sample. BMC Med Res Methodol 5:1–10. https://doi.org/10.1186/1471-2288-5-13/TABLES/3

Hubka V, Eder W (2001) Functions revised. In: International conference on engineering design (ICED 2001). Glasgow, U.K., Aug. 21–23, 2001

Ignacio P (2022) Can engineers be primed to think in systems? An empirical study showing the effects of concept mapping on engineering students' ability to explore the design space.

Kroll E (2013) Design theory and conceptual design: Contrasting functional decomposition and morphology with parameter analysis. Res Eng Des 24:165–183. https://doi.org/10.1007/S00163-012-0149-6/FIGURES/17

Krüger MF, Zorn S, Gericke K (2023) Combining function modelling and requirements modelling with the ifm framework. In: Proceedings of the design society. Cambridge University Press (CUP), pp 987–996

Lee Y, Ho FN, Wu MC (2018) How do form and functional newness affect adoption preference? The moderating role of consumer need

for uniqueness. J Consum Mark 35:79–90. https://doi.org/10.1108/JCM-10-2015-1578/FULL/PDF
Majumder A, Todeti SR, Chakrabarti A (2023) Empirical studies on conceptual design synthesis of multiple-state mechanical devices. Res Eng Des 34:477–495. https://doi.org/10.1007/S00163-023-00420-8/FIGURES/13
Malmqvist J (1995) A computer-based approach towards including design history information in product models and function-means trees. Proc ASME Des Eng Tech Conf 2:593–602. https://doi.org/10.1115/DETC1995-0193
Manikandan S (2011a) Measures of central tendency: median and mode. J Pharmacol Pharmacother 2:214. https://doi.org/10.4103/0976-500X.83300
Manikandan S (2011b) Measures of central tendency: the mean. J Pharmacol Pharmacother 2:140. https://doi.org/10.4103/0976-500X.81920
Mathieson JL, Shanthakumar A, Sen C et al (2012) Complexity as a surrogate mapping between function models and market value. Proc ASME Des Eng Tech Conf 9:55–64. https://doi.org/10.1115/DETC2011-47481
McAdams DA, Stone RB, Wood KL (1999) Functional interdependence and product similarity based on customer needs. Res Eng Des 11:1–19. https://doi.org/10.1007/S001630050001
Mesa JA (2023) Design for circularity and durability: an integrated approach from DFX guidelines. Res Eng Des 34:443–460. https://doi.org/10.1007/S00163-023-00419-1/TABLES/9
Mesa JA, González-Quiroga A (2023) Development of a diagnostic tool for product circularity: a redesign approach. Res Eng Des 34:401–420. https://doi.org/10.1007/S00163-023-00415-5/TABLES/11
Nagel RL, Bohm MR (2011) On teaching functionality and functional modeling in an engineering curriculum. In: Proceedings of the ASME design engineering technical conference. Washington, DC, Aug. 28–31, pp 625–636
Nagel RL, Bohm MR, Linsey JS, Riggs MK (2015) Improving students' functional modeling skills: a modeling approach and a scoring rubric. J Mech Des Trans ASME 137:051102. https://doi.org/10.1115/1.4029585
Narum SR (2023) Correction: beyond bonferroni: less conservative analyses for conservation genetics (Conservation Genetics, (2006), 7, 5, (783–787), https://doi.org/10.1007/s10592-005-9056-y). Conserv Genet. https://doi.org/10.1007/S10592-023-01576-5/METRICS
Pahl G, Beitz W, Feldhusen J, Grote KH (2007) Engineering design: a systematic approach (3rd edition), 3rd edn. Springer, London
Patel A, Kramer WS, Flynn M et al (2020) Function modeling: a modeling behavior analysis of pause patterns. J Mech Des 142:111402. https://doi.org/10.1115/1.4046999
Patel A, Kramer WS, Flynn M et al (2017) Function modeling: comparison of chaining methods using protocol study and designer study. In: International design engineering technical conferences & computers and information in engineering conference. ASME, Cleveland, Ohio
Perneger TV (1998) What's wrong with Bonferroni adjustments. BMJ 316:1236–1238. https://doi.org/10.1136/BMJ.316.7139.1236
Reeling H, She J (2023) Aligning functional analysis processes with designers' natural cognitive flow. In: The 24th international conference on engineering design. Bordeaux, France, Jul. 24–28, 2023
Robotham AJ (2010) The use of function/means trees for modelling technical, semantic and business functions. J Eng Des 13:243–251. https://doi.org/10.1080/09544820110108944
Schön DA (2017) The reflective practitioner: How professionals think in action. The Reflective Practitioner: How Professionals Think in Action 1–374. https://doi.org/10.4324/9781315237473/REFLECTIVE-PRACTITIONER-DONALD-SCH
Sen C, Caldwell BW, Summers JD, Mocko GM (2010a) Evaluation of the functional basis using an information theoretic approach. AI EDAM 24:87–105. https://doi.org/10.1017/S0890060409990187
Sen C, Summers JD, Mocko GM (2010b) Topological information content and expressiveness of function models in mechanical design. J Comput Inf Sci Eng. https://doi.org/10.1115/1.3462918/464128
Shah JJ, Kulkarni SV, Vargas-Hernandez N (2000) Evaluation of idea generation methods for conceptual design: effectiveness metrics and design of experiments. J Mech Des 122:377–384. https://doi.org/10.1115/1.1315592
Shah JJ, Vargas-Hernandez N, Summers JD, Kulkarni S (2001) Collaborative sketching (C-Sketch)—an idea generation technique for engineering design. J Creat Behav 35:168–198
Shankar P, Morkos B, Yadav D, Summers JD (2020) Towards the formalization of non-functional requirements in conceptual design. Res Eng Des 31:449–469. https://doi.org/10.1007/s00163-020-00345-6
She J, Belanger E, Bartels C, Reeling H (2022) Improve syntax correctness and breadth of design space exploration in functional analysis. ASME J Mech Des 144:111402. https://doi.org/10.1115/1.4054875
Simon HA (1973) The structure of ill structured problems. Artif Intell 4:181–201. https://doi.org/10.1016/0004-3702(73)90011-8
Summers JD, Eckert C, Goel AK (2017) Function in engineering: benchmarking representations and models. Artif Intell Eng Des Anal Manuf AIEDAM 31:401–412. https://doi.org/10.1017/S0890060417000476
Tomko M, Nelson J, Nagel RL et al (2017) A bridge to systems thinking in engineering design: an examination of students' ability to identify functions at varying levels of abstraction. AI EDAM 31:535–549. https://doi.org/10.1017/S0890060417000439
Ullman DG (2017) The mechanical design process, 6th edn. David Ullman LLC
Ulrich KT, Eppinger SD, Yang MC (2020) Product design and development (seventh edition). McGraw Hill
Van Eck D, Weber E (2021) Assessing function modeling frameworks: technical advantage predictions as a conceptual tool. Eng Stud 13:205–225. https://doi.org/10.1080/19378629.2021.1989441
Wong FS, Wynn DC (2023) A systematic approach for product modelling and function integration to support adaptive redesign of product variants. Res Eng Des 34:153–177. https://doi.org/10.1007/S00163-022-00401-3/FIGURES/15
Yildirim U, Campean F (2020) Functional modelling of complex multi-disciplinary systems using the enhanced sequence diagram. Res Eng Des 31:429–448. https://doi.org/10.1007/S00163-020-00343-8