Do human and computational evaluations of similarity align? An empirical study of product function

Ananya Nandy
Dept. of Mechanical Engineering
University of California, Berkeley
Berkeley, California
Email: ananyan@berkeley.edu

Kosa Goucher-Lambert [*]
Dept. of Mechanical Engineering
University of California, Berkeley
Berkeley, California
Email: kosa@berkeley.edu

[*] Address all correspondence to this author.

ABSTRACT

Function drives many early design considerations in product development, highlighting the importance of finding functionally similar examples if searching for sources of inspiration or evaluating designs against existing technology. However, it is difficult to capture what people consider is functionally similar and therefore, if measures that quantify and compare function using the products themselves are meaningful. In this work, human evaluations of similarity are compared to computationally determined values, shedding light on how quantitative measures align with human perceptions of functional similarity. Human perception of functional similarity is considered at two levels of abstraction: (1) the high-level purpose of a product, and (2) how the product works. These human similarity evaluations are quantified by crowdsourcing 1360 triplet ratings at each functional abstraction and creating low-dimensional embeddings from the triplets. The triplets and embeddings are then compared to similarities that are computed between functional models using six representative measures, including both matching measures (e.g. cosine similarity) and network-based measures (e.g. spectral distance). The outcomes demonstrate how levels of abstraction and the fuzzy line between "highly similar" and "somewhat similar" products may impact human functional similarity representations and their subsequent alignment with computed similarity. The results inform how functional similarity can be leveraged by designers, with applications in creativity support tools, such as those used for design-by-analogy, or other computational methods in design that incorporate product function.

1 INTRODUCTION

Designers often make comparisons between different ideas and assess how their designs will meet functional requirements to solve the problem at hand. To acquire knowledge in early-stage design, a common practice is to seek examples of, or inspiration from, existing products, through methods such as benchmarking or searching patents [1,2]. Previous work has shown that inspirational stimuli help improve idea generation and that function-based examples are particularly useful in helping designers identify potential solutions [3–5]. Several quantitative approaches have been applied to determine functional similarity between products, guiding the development of computational methods to augment designers' capabilities in the solution exploration phase [6–8]. However, the alignment of these methods with how designers draw functional connections between products in practice (and consequently, their utility), remains to be understood. Depending on the stage of the design process, designers may consider concepts at different levels of abstraction [9, 10]. While functional representations are abstractions themselves, they may vary in level of detail. Some consist of only the core function of the product, while others consist of all of the sub-functions that make the product work. When searching for concepts across domains, prioritization may lie on the higher-level function of the product (here, referred to as its purpose) to find surface dissimilar ideas. At other times, finding products with similar functional properties might entail searching for the specific mechanisms necessary to achieve this purpose (i.e., how it works) [11–14]. When considering function in these different ways, products that are relevant for benchmarking, as examples, or for inspiration, may vary, motivating the need for appropriate similarity measures. While humans may be able to adapt their consideration of functional similarity across these abstractions, automated methods must contend with this often ambiguous notion of similarity.

To investigate how design similarity can be assessed and utilized for the early stages of design, the following research questions are addressed in this work:

1. Do computed measures of functional similarity accurately capture human representations of functional similarity?
2. How does the level of abstraction impact humans' similarity representations of product function?

Aiming to ensure that computational methods to support design are meaningful to humans, a quantitative approach is taken to compare human conceptualization of similarity with how similarity can be measured mathematically. Crowdsourcing methods are applied to quantify human-determined functional similarity and explore how various quantitative measures align with the human representations. While the measures included for comparison are not comprehensive, they are representative of measures that have been investigated in prior work or may be relevant for engineering design. The results show how well measures of functional similarity match humans' perceptions of similarity and in which cases (e.g. when considering only highly similar designs vs. a range of products or for higher vs. lower functional abstraction), leading to broader considerations of how human vs. computational representations of functional similarity might be applicable for computational design tools.

2 RELATED WORK

Prior research on the use of similarity measures in engineering design, as well as research on human perception of similarity and its measurement are reviewed in the following section. Both of these areas are relevant for comparing human evaluations with computational output, ensuring that the latter is interpretable and useful to humans during design activities.

2.1 Similarity measures used in design

Obtaining repeated evaluations of design similarity from humans, through expert or crowd-sourced assessments, is challenging and expensive, prompting effort towards finding quantitative similarity measures for the design domain. Since design spans various tasks and contexts, it is often desirable to adapt existing measures that have already been shown to apply across various domains. Similarity has been assessed on different dimensions, such as form or function, and at different phases, ranging from concepts to full products [15]. For example, visual similarity between products (similarity in form) has been investigated for the purpose of determining product families, variants, or branding [16, 17]. In the early stages of design, however, product function is often one of the most critical considerations [2]. Assessing similarities along the dimension of product function poses a challenge because product function is difficult to quantify.

One way to calculate functional similarity involves using a text-based repository of domain relevant information such as the patent database, which contains a large body of data on product function. Functional similarity has been calculated using latent semantic analysis and the cosine similarity measure on these patents for the purpose of design-by-analogy (a method where designers seek to apply solutions that work for other problems to solve their problem) [6]. The results from using this measure has been validated by indicating that its clustering of patents is sensible to experts [18]. A functional neural network has also been used for presenting function-based inspirational stimuli by defining product function via relationships between parts and neighboring parts within 3D models [19]. Another way to capture product function is through a functional diagram or model, such as one developed using a standardized vocabulary [2,20,21]. A vector-based quantitative metric has been developed to compare these functional models [7]. In addition, critical function chains have been extracted from functional models and matched in various ways to quantify functional similarity [8, 22]. The functional model representation enables a higher level of abstraction of a product than a patent, which may be desirable when searching for examples during conceptual design. At the same time, functional models are not available for many products and are often developed subjectively. To mitigate these challenges, recent research has focused on automating functional modeling using information about product components [23, 24]. This research focuses on functional similarity as obtained through functional models. Similarity can be calculated in a variety of ways across these models, yet it remains unclear whether these determinations of functional similarity are meaningful. It is also notable that each measure provides a different conceptualization of what similarity means when applied to product function. Some measures place importance on the existence (or absence) of specific sub-functions to define overall functional similarity, while others place more importance on patterns in how sub-functions connect to each other. These differences impact the context in which each measure can be used and indicates the necessity to carefully consider how functional similarity is quantified, especially when it is used to support design activities [25].

2.2 Quantifying how humans evaluate similarity

To assess how quantitative measures align with human mental representations, it is crucial to capture how humans perceive similarity. Knowing the “human dimension” is important when applying these measures within interventions or systems that are intended to augment human processes. Humans constantly make judgements of similarity to reconcile information from the world around them with internal mental representations. In addition, similarity is said to play a part in how people structure conceptual knowledge [26]. Several theories have been developed regarding human determination of similarity, incorporating effects from elements such as directionality and context [27, 28]. In the structural alignment view of similarity from psychology, there are three elements of alignment: structural consistency, relational focus, and systematicity. These elements correspond to one-to-one matching, common relations in both items being compared, and sets of relations that are interconnected by higher order relations [29]. It is difficult to untangle the underlying dimensions along which people consider similarity, especially for more complex items, and there are several approaches to tackling this challenge.

One approach to understanding why people might consider two objects to be similar is to explicitly ask them for their reasoning. For example, to evaluate how a topic modeling algorithm represents similarity compared to humans, human raters were asked to both select which documents were more similar from a triplet as well as explain what made them similar (and the unchosen one different) [30]. Another approach is a data-driven approach, where people are asked to make similarity judgements and latent or explainable dimensions are uncovered directly from the results. This data-driven approach has been used in several contexts such as determining similarity across musical artists and natural objects [26, 31]. Closely related to the approach in our work, the data-driven approach has been utilized to create an embedding to compare human-perceived similarities with models' internal representations on ImageNet [32].

Within engineering design, both approaches have been used to assess design similarity for a variety of purposes. In a study on design-by-analogy, participants were explicitly asked what dimensions they considered important for similarity between a target and source product [33]. It was found that functional similarity dominated over form similarity. To understand whether the structural alignment view of similarity from psychology applies in the context of design, participants were asked to rate similarity between design concepts and explain their reasoning in another study [34]. The results implied that feature-based responses drove similarity, in line with the element of structural consistency from the structural alignment model.

More recently, the data-driven approach has been increasingly applied to problems in design. Pairwise similarity judgements were crowdsourced to assess visual similarity between products, determining that novelty assessments from a crowd can match with those made by experts [35]. Similarity judgements were also collected in the form of triplet ratings for determining design sketch novelty and for evaluating dissimilarity between sets of ideas to spur diversity during idea generation [36, 37]. More fine-grained search across product function specifically has been enabled by crowdsourcing annotations from product descriptions, including both product purpose and working mechanism as facets [14].

These data-driven methods are able to uncover human perceptions of similarity, but may be limited to the task or context for which the data were collected. In addition, the dimensions of similarity determined from data-driven methods may not be explanatory or easy to interpret. Therefore, using similarity functions that have been learned from humans may not always be possible or desirable. At the same time, even a similarity measure that is computed from products must provide human-interpretable results to successfully supplement cognitive processes such as analogical transfer. For these reasons, the determination of functional similarity is approached here in a task-agnostic — though not domain-agnostic — way, combining a data-driven approach (collecting human similarity judgements) with one that is less context-dependent (using mathematical measures on functional diagrams).

2.3 Considerations for evaluating design similarity

Several factors may influence whether and how quantitative similarity measures, computed on designs, can be effectively used to support human creativity in engineering design. Two are considered here: the threshold for separating similar and dissimilar and the consideration of ideas at varying abstraction levels during design.

The concepts of similarity, dissimilarity, and distance are often used interchangeably. Mathematically, distance measures can be converted to dissimilarity measures. In addition, similarity can be converted to dissimilarity, and vice versa. However, in application, whether similarity or dissimilarity is more important depends highly on the context. For example, in recommendation systems broadly, while similarity is used to find the most relevant results, the notion of dissimilarity is explored instead to add novelty and diversity to results [38]. In the domain of engineering design, dissimilarity has been applied to developing novelty metrics for idea assessment at the conceptual stage. To characterize novelty, these metrics emphasize how dissimilar a given concept is from other concepts [39].

Measures of similarity also play a critical role in attempts to foster analogical innovation. According to Gentner and Markman, the processes involved in comparisons of similarity and analogy are the same [29]. Analogical distance has been shown to impact the effectiveness of examples during concept generation, indicating that the distinction between similarity and dissimilarity is critical in the practice of design-by-analogy [40–42]. Analogies are generally considered to be near-field analogs (sharing surface features or existing in the same domain) or far-field analogs (sharing few or no surface features and existing in different domains, but having some functional similarity). As such, analogical distance encompasses similarity and dissimilarity as well as balancing the line between the two and not going “too far” [43]. Functional similarity may be utilized in design to find both items that are highly similar and ones that are “somewhat similar.” Therefore, in this research, comparisons of measures and human judgements are considered with respect to both highly similar products and similarity in the product space more globally.

Another element to consider when assessing similarity between designs or products is the abstraction level of their representations. Work in the cognitive processes behind design suggests that solution search is performed through lateral and vertical transformation: moving to a slightly different idea or moving to a more detailed version of the same idea [44]. In the context of product function, as the level of detail available increases, the functional abstraction can decrease, facilitating consideration of function as how the product works instead of its higher-level purpose [12]. Functional similarity measures have focused on lower-level representations (i.e. the working mechanism) since very detailed information is available in patents and in full functional models that have been developed through reverse engineering. However, designers often only have enough information to operate at the higher level during the conceptual stages of design. In addition, cross-domain analogies can be found through higher-level purpose even if working mechanisms differ [13]. Because it is often required to consider functional similarity at multiple levels of abstraction, it is necessary to understand how any quantitative measures reflect the ways humans can translate between the levels. This work encodes the dimension of functional abstraction to specifically examine its influence on human representations of similarity.

3 METHODS

Functional similarity is crowdsourced from humans and compared with similarity recovered from applying quantitative measures directly on the products, as outlined in the following sections. The comparisons can provide insight into how example-based design tools or computational methods might capture product functionality, making this information accessible to designers.

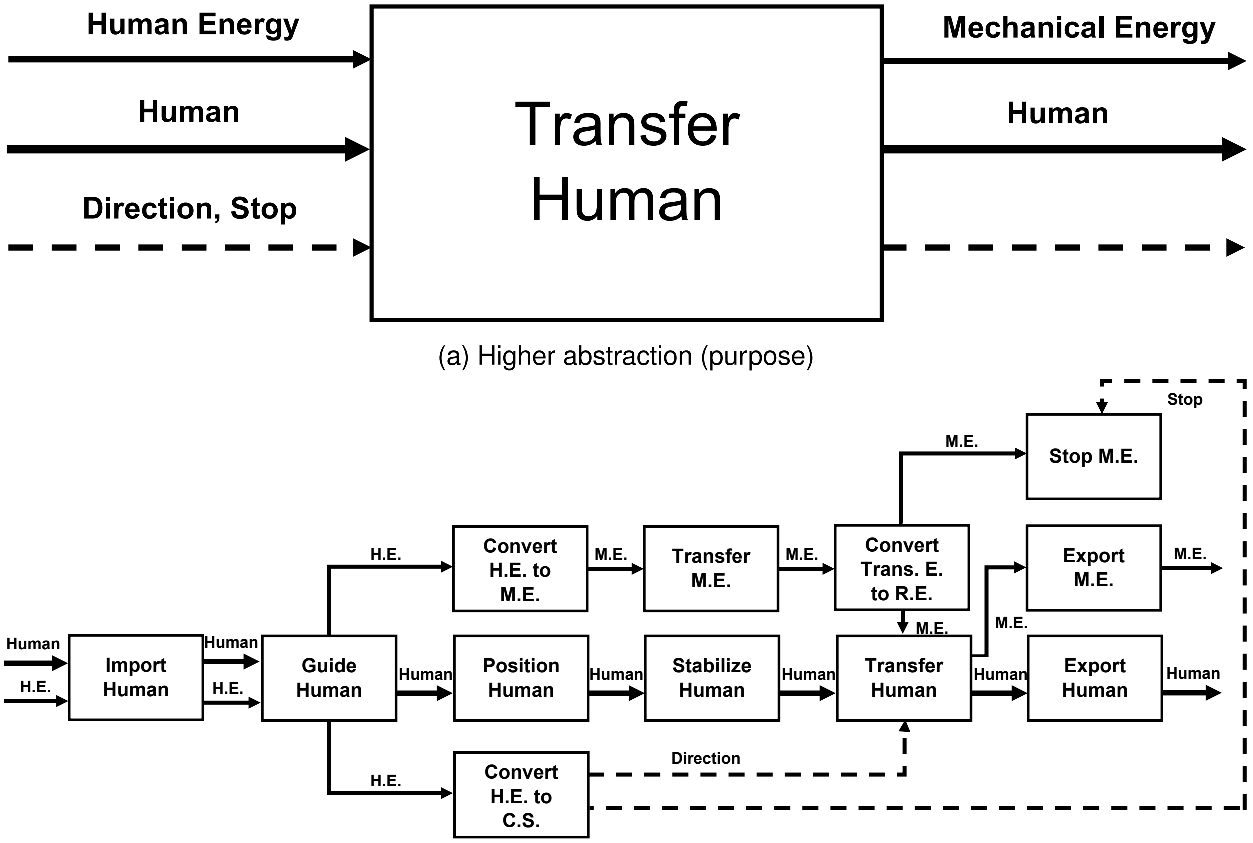
3.1 Product dataset

A subset of 20 consumer products (e.g. toys, consumer electronics, household devices) found in the Design Repository hosted by Oregon State University was utilized for this work [45]. This subset was selected to represent products with varying levels of complexity that participants would be familiar with, as well as to ensure the availability of two consistent levels of functional specification. A list of the products can be found in Appendix A. For each product, the repository contained a simple functional model consisting of inputs, outputs, and a singular, main function of the product, as well as a highly detailed functional model of how the product worked, specified according to Hirtz et al. [21]. An example of each type of functional model is shown in Figure 1. The repository additionally contained a product title and image.

3.2 Crowdsourcing human judgements

To capture how humans consider products to be functionally similar, similarity judgements were crowdsourced from humans and then represented in a low-dimensional embedding space using techniques from machine learning. These embeddings were used to quantify the relative similarity among the set of products. This method has recently been used in engineering design to determine the visual similarity of products as well as to determine the novelty of ideas [35, 36]. The judgements were collected in the form of triplet queries ("Is A more similar to B or to C?"). Prior work has shown that humans can more easily and consistently answer triplets as opposed to direct pairwise comparisons [46].

Because functional similarity may depend on the level of abstraction at which someone is considering products, information about function was presented to participants in two ways based on the two available types of functional models, shown in Figure 1. The information from the full functional models (Figure 1b) was converted to text descriptions to capture essential information



(a) Higher abstraction (purpose)

(b) Lower abstraction (working mechanism)

Fig. 1: Functional models at two levels of abstraction (shown for a scooter)

about how the product worked. The function as defined by its purpose was taken directly as text from the simple functional model (Figure 1a) and only modified in a few cases for clarity. The descriptions can be found in Appendix A. Stock images were included for products that were missing images in the Design Repository and product titles were modified to represent the generic versions of the product. Each triplet presented to participants contained the following information about each product: a title, image, and description of the product function of either type. An example triplet is shown in Figure 2.

Although images were provided to aid participants in understanding what the products were, they were instructed to judge similarity along the dimension of function and not form. The participants were instructed to consider the overall purpose of the product when presented with the shorter descriptions and to consider the way in which the product worked when presented with the

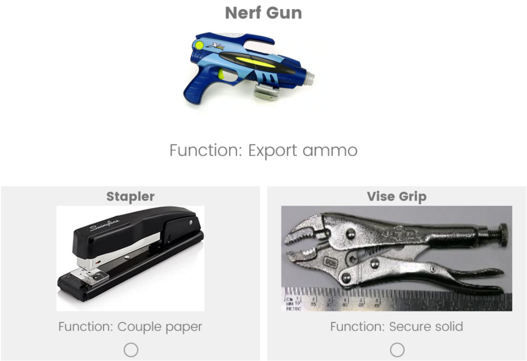


Fig. 2: Example of a triplet query. The function description text displayed under the three images were descriptions from one of the two levels of abstraction (detailed in Appendix A). Here, the higher abstraction is shown.

longer description. Each participant was provided with a subset of triplets (randomized across all possible triplets) and made judgements on this same subset of triplets twice, once presented with the shorter descriptions and once presented with the longer description. The order of the triplets was randomized across the two abstraction levels.

After approval from an Institutional Review Board, data was collected from a total of 69 participants. Data from one participant was removed, as they did not follow instructions. The included participants consisted of 42 undergraduate students, 16 graduate students, and 10 others (including working professionals). Among the participants were 50 who identified as male, 17 who identified as female, and 1 who preferred not to say. A majority of participants were pursuing, or had graduated with, a mechanical engineering degree, indicating a level of domain expertise. 24 of the participants indicated that they had greater than 4 years of engineering/design experience through courses, work, or extracurricular activities. 36 of the participants were shown the longer, lower-abstraction descriptions first, while 32 were shown the shorter, higher-abstraction descriptions first.

A total of 2720 triplet ratings were collected from the participants, who each provided 40 ratings. Half of each set of ratings (1360 triplets) contained information about each level of functional abstraction. Therefore, each set of triplets collected consisted of 42 percent of the total possible triplets (3240 triplets). Previous applications of the low-dimensional embedding techniques considered here use 20 percent of triplets [47, 48]. Additionally, prior work that incorporates the full triplet set found that using 30 percent of the triplet set was sufficient and robust to a small number of false ratings [36].

3.3 Generating an embedding space for human judgements

Once the triplets were collected, they were mapped to a low-dimensional space. The data embeddings can be constructed by using the triplets as constraints to where points (in this case, products) are placed within the n-dimensional space. One machine learning method to do this is t-Distributed Stochastic Triplet Embedding (t-STE). This method defines a probability density distribution (a heavy-tailed kernel) and maximizes these probabilities with respect to the embedding points so that a triplet is satisfied. Additionally, the maximization ensures that similar points are collapsed while dissimilar points kept apart by triplets are repelled [48]. Other commonly used embedding methods include Generalized Non-metric Multidimensional Scaling (GNMDS), Crowd Kernel Learning (CKL), and Stochastic Triplet Embedding (STE) [46–48]. GNMDS and t-STE have been explicitly applied to work in the design domain [36, 37]. In this work, t-STE was chosen as the embedding method, due to its ability to ensure similar points are closer together and dissimilar points are farther apart, without violating any constraints [48]. Preliminary analyses demonstrated that selecting t-STE as the embedding method (as opposed to one of the aforementioned techniques) did not significantly affect the results.

To address partial triplet collection and the aggregation of triplet ratings from across the population of participants, two measures, adapted from prior work, were used to determine the quality of the embedding: distance error and triplet generalization error. Distance error refers to the mean squared error between the the normalized Euclidean distances derived from the final embedding and an embedding created with consecutively fewer triplets [36]. This measure was used to determine how much the embedding changes with the addition of new triplets to ensure that there are enough triplets for convergence. Triplet generalization error is calculated by holding out a set of triplets, calculating the embedding, and then determining whether the calculated embedding satisfies the triplets that were held out [48]. This measure was used to assess how successfully the methods could satisfy triplets that were not provided.

Finally, the dimensionality hyperparameter of the t-STE embedding was considered when comparing to computed measures in case two dimensions did not best represent the human-determined functional similarity embedding (though all of the embedding techniques have been used to represent human judgements from other domains in two dimensions). This hyperparameter was alternately set to thirteen, found by creating the embedding with an increasing number of dimensions and determining the minimum number of dimensions at which the embedding's cost value converged. Once the embedding was created from triplet ratings, the Euclidean distances between the points were calculated, range normalized, and converted to a pairwise similarity matrix.

3.4 Measuring similarity directly from products

3.4.1 Computing pairwise similarity matrices from functional models

After the human conceptualization of function was quantified, the next step was to compare this to how quantitative measures determined functional similarity. To do this, the full, lower-level functional models (Figure 1b) for the same set of products were represented in a mathematical space as binary matrices, specified using 21 functions and 19 flows (as defined by the functional basis framework). Within this framework, a 1 was used for the existence of a specific function in the product and a 0 was used for the absence of a specific function in the product [21].

The quantitative measures of similarity used were those considered extensively in prior work by the authors [25]: simple matching coefficient, Jaccard similarity, cosine similarity, spectral distance [49], NetSimile [50], and DeltaCon [51]. There are many possible ways to measure similarity and not all were included here, but the six measures represent different characterizations of similarity when applied to functional representations of products. The measures range from those that are easily interpretable to those that are not. In addition, although the measures represent more general formulations and have been applied across several domains, efforts were made to select measures that were the most meaningful for the context of engineering design. For example, versions of cosine similarity and a matching measure much like the simple matching coefficient or Jaccard similarity have been applied to engineering design [6–8]. More details on these measures specifically applied to functional models can be found in our previous work [25] as well as in the summary presented in Table 1. Methods requiring significant training data such as neural networks were not considered, though similarity embeddings from such networks could be compared to the "human" embedding generated in this work if similar products are included.

The SMC, Jaccard, and cosine measures involve variations of matching the existence of features (in this case, functions or flows) across the products being compared. The spectral, NetSimile, and DeltaCon measures involve modeling the products as networks and then comparing the network structure in various ways. For example, the spectral measure incorporates information node degree, which refers to the number of sub-functions operating on a specific flow or a sub-function operating on a number of flows. This could represent the relative importance of specific functions and flows within a functional model. The two different types of measures, represent possible links to *one-to-one matching* and *relational comparison*, both aspects of the structural alignment model of how humans determine similarity [29].

Table 1: Summary of similarity measures used for comparison against human judgements. A description of how the measure works and the general measure type is provided, though further details on the use of these measures in the context of calculating similarity between functional models can be found in our prior work [25].

Measure	Definition	Type
SMC	The intersection over the union of sample sets, including mutual absences and presences	Matching
Jaccard	The intersection over the union of sample sets with only mutual presences	Matching
Cosine	Normalized dot product of vectors	Matching
Spectral	Distance between normalized Laplacian (degree matrix minus adjacency matrix) of graphs	Network
NetSimile	Distance between aggregated feature (e.g. clustering coefficient, node neighbors, etc.) vectors of graphs	Network
DeltaCon	Differences in corresponding node affinities (influence of one node on another) of graphs	Network

The measures were calculated on the product function matrices using the SciPy, NetworkX [52], and NetComp [49] libraries in Python to obtain pairwise comparisons. These pairwise comparisons were range normalized and converted to similarities if the original form was a distance or dissimilarity. Therefore, all relative comparisons were scores between 0 and 1, with 1 representing the highest similarity (only for a product compared to itself) and 0 representing the lowest similarity.

3.4.2 Generating triplets from computed similarity

The pairwise similarities calculated using functional models were converted into triplet form for direct comparison with the triplets collected from participants. Triplets were generated from the pairwise comparison matrices from any of the similarity measures. Given products A, B, and C, if the pairwise similarity of A and B was greater than that of A and C, the generated triplet for the triplet query (”Is A more similar to B or to C?”) was “A is more similar to B than to C.” These generated triplets were found for all possible triplet combinations since the full pairwise matrix is available from computed similarities. If there were ties between the similarity of A and B and A and C, as is possible with measures such as the Jaccard measure, the more similar product was selected randomly.

3.5 Comparing human judgements with computed similarity matrices

Creating an embedding from the crowdsourced data allows comparisons of functional similarity at the level of specific products without the need to collect the full set of triplets. This is important since the complete set of triplets scales up rapidly as the number of products increases. However, to address the noise introduced to the human evaluations by the process of learning the embedding, the original crowdsourced triplets were used as another representation of the human perception of functional similarity. Thus, the crowdsourced and computational results were compared in the following two ways to investigate the agreement between human and computational similarity assessment: (1) by using pairwise similarity matrices from the human embedding vs. pairwise similarity matrices computed from similarity measures and (2) by using the triplets collected from humans vs. triplets induced by computed similarity matrices.

3.5.1 Comparison of full embeddings via correlation of pairwise similarity matrices

After the creation of a human embedding, functional similarity was assessed relative to all other products in the considered set of products. Typically, these relative rankings are compared using correlation statistics, such as Kendall’s τ (useful for rankings with ties) or Spearman’s ρ, as has been done in representational similarity analysis [53]. To determine an overall comparison between each human embedding and the results from each similarity measure, Kendall's τ was computed between the upper triangular matrices of the two pairwise matrices (human and computed). This rank correlation is interpreted as the difference in probability that ranks are in the same order vs. in different order, across the rankings being compared.

3.5.2 Matching triplet ratings

Matching the triplets collected from participants with the triplets generated from each similarity measure allows comparison without necessarily relying on the quality of the learned embedding. This matching was determined by using the intersection between each participant's answered triplets – a subset of all possible triplets – and the generated triplets (a triplet only matches if the exact order is the same), divided by the number of triplet queries answered by the participant (20, for this study). Therefore, the measure allows for assessment of how well quantitative similarity assessments are matching humans generally, as well as between specific measures and abstraction levels.

3.5.3 Comparison of product-level functional similarity via normalized discounted cumulative gain

Finally, to look more closely at the level of individual products, the comparison of human and computational output was formulated as a search problem: a product was selected as if it was the input of a search and all other products were ranked relative to that product. Normalized discounted cumulative gain (NDCG), often used to assess recommender systems, was adapted from the field of information retrieval [54]. NDCG can be used to compare rankings to a "ground truth," given relevance scores, with the higher ranks having more importance than lower ranks.

The discounted cumulative gain (DCG) can be found by using a logarithmic discount based on the rank position (i is the rank position, rel_i is the relevance at rank position i, and n is the total number of ranks) as following

$$DCG = \sum_{i=1}^{n} \frac{rel_i}{log_2(i+1)}, \tag{1a}$$

after which it must be normalized by the ideal discounted cumulative gain (IDCG). The IDCG refers to the value of DCG when the list is sorted in order of relevance so that the highest rank has the highest relevance [54].

$$nDCG = \frac{DCG}{IDCG} \tag{1b}$$

In this case, the crowdsourced human rankings were considered to be the "ground truth" and each numerical rank from the crowdsourced ranking was used as a relevance score. This was used to calculated the IDCG. Then, the DCG was calculated for the ranking of the , to compare between the human and computational measure (e.g. cosine similarity) being compared. Once again, the numerical ranks were used as a relevance score.

Furthermore, when using relative pairwise comparisons, the comparisons must be made using rankings instead of absolute scores since the distribution of values generated across the different similarity measures varies [25]. However, converting to rankings leads to loss of information about whether a product or set of products in the ranking are significantly farther away overall (i.e. the global structure of the product space). In other words, the individual rankings might include products that have very little or no relevance to each other. To probe whether alignment between measures and humans is driven by products that are considered highly similar or by the overall similarity space, thresholds were explored to try to separate relevant products from non-relevant products within the entire product space. The thresholds were based on similarity above a percentile, using the entire pairwise score matrix and products below the threshold would be given a relevance of 0 when appearing in any ranking. At the individual product level, a product could have as few as 0 products or as many as 9 other products considered to be relevant in its ranking. If the rankings by humans and a quantitative measure were exactly the same, the NDCG would return a value of 1. NDCG was calculated using the Scikit-learn library in Python. Figure 3 summarizes the steps taken to compare human evaluations with the calculated similarities.

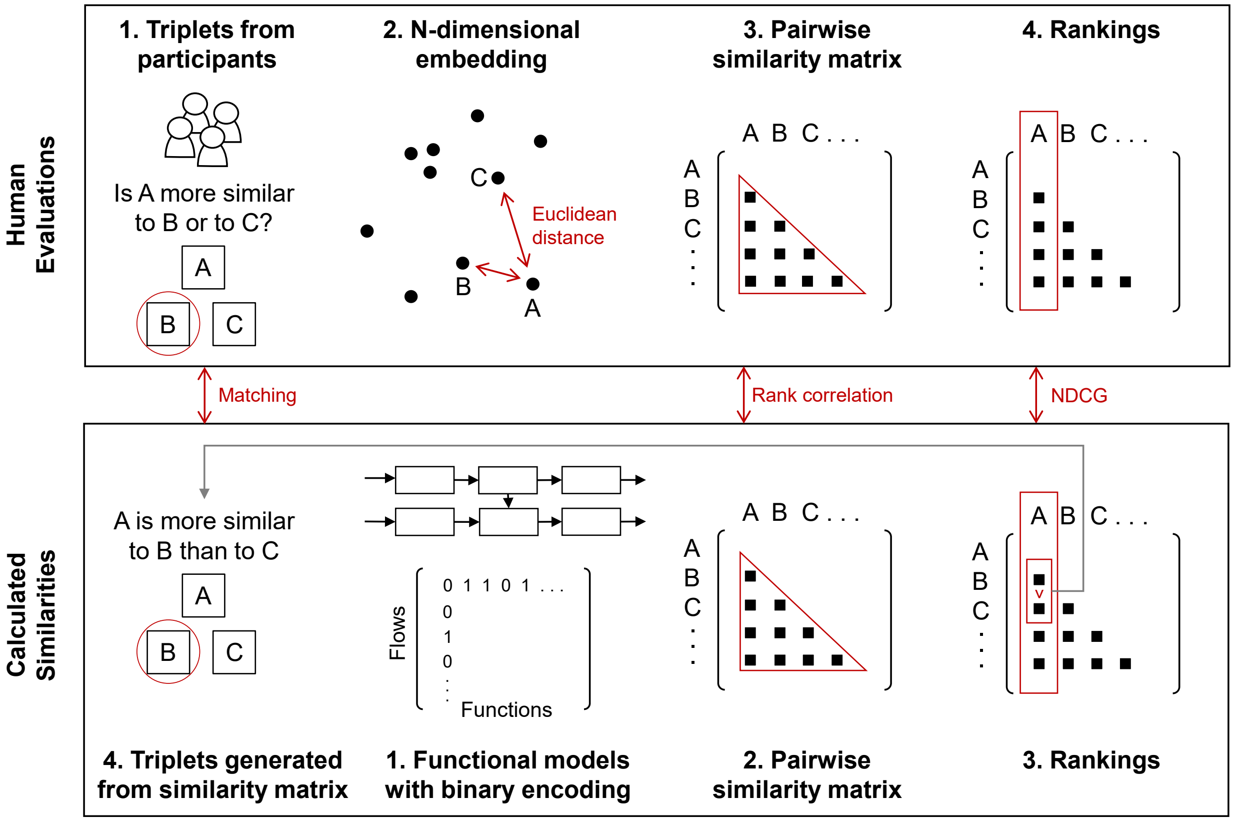

Human Evaluations
1. Triplets from participants
Is A more similar to B or to C?
A
B
C
2. N-dimensional embedding
C
B
A
Euclidean distance
3. Pairwise similarity matrix
A B C . . .
A B C . . .
4. Rankings
A B C . . .
A B C . . .
Matching
Rank correlation
NDCG
Calculated Similarities
A is more similar to B than to C
A
B
C
0 1 1 0 1 . . .
0
1
0
Flows
Functions
A B C . . .
A B C . . .
>
4. Triplets generated from similarity matrix
1. Functional models with binary encoding
2. Pairwise similarity matrix
3. Rankings

Fig. 3: Sequence of steps taken to compare human evaluations of functional similarity with calculated functional similarity. The comparisons are made in two ways: using pairwise similarity matrices and rankings (steps 3-4 and 2-3 respectively) and using triplets (steps 1 and 4 respectively).

4 RESULTS

Similarity determined from crowdsourced human data is compared to the calculated similarity scores using the methods outlined in Section 3 and considering levels of abstraction. First, the human representation of product function similarity is quantified at both the higher- and lower-level abstraction (i.e. the product's purpose vs. the product's working mechanism) into a low-dimensional embedding space through t-STE. Other methods for creating the embedding are compared to verify the quality of using the t-STE embedding method. Next, both the human embeddings and raw triplets are compared across the lower and higher abstraction to investigate the effect of the functional abstraction on participants' representation of functional similarity. Then, the crowdsourced and computational results are compared in both forms, across abstraction levels. Finally, more qualitative investigations are conducted at the level of individual products to determine if differences appear based on the context of use for the similarity measure.

4.1 Quantifying human perception of functional similarity in an embedding space

The collected triplets at each abstraction level are used to create a low-dimension embedding of the product space using t-STE. Figure 4 provides a visualization of which products were considered functionally similar by participants under the perspective of function as a product's working mechanism in 2D. A similar embedding is created for the higher abstraction level (function as a product's purpose). Before using the pairwise similarity matrix derived from Figure 4 in further analysis, some steps are taken to ensure that the generated embedding provided a satisfactory representation of the human data.

Creation of the embedding is replicated using the three other common triplet embedding methods (GNMDS, CKL, and STE). For all of the methods, triplet generalization error and distance error are calculated using fractions of the collected triplets to the full number of collected triplets. As shown in Figure 5a, the GNMDS, STE, and t-STE methods demonstrate a level of convergence before the full number of collected triplets are included. The t-STE method has the lowest triplet generalization error by a small margin when incorporating all of the collected triplets. Even using the full number of collected triplets, about 30 percent of the triplet constraints are not satisfied in the embedding. The occurrence of unsatisfied constraints is in line with previous experiments us-

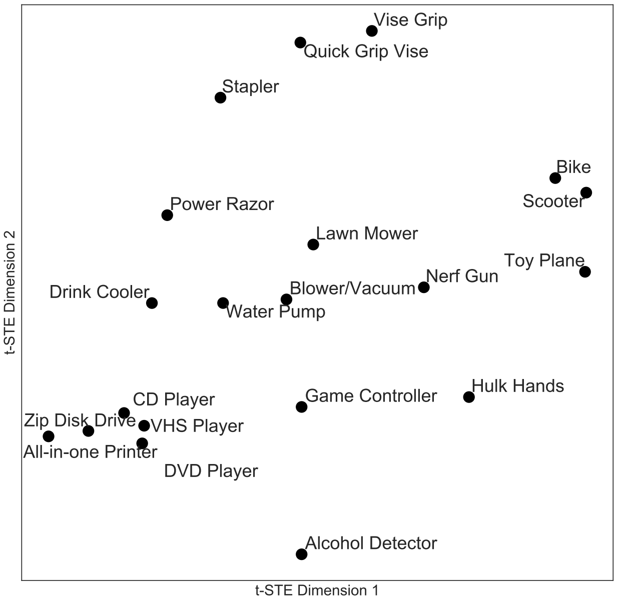
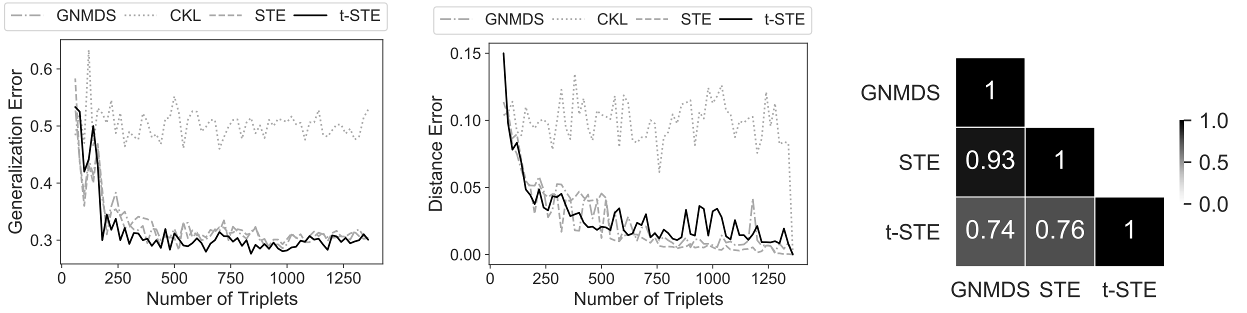


Fig. 4: Two-dimensional embedding constructed with t-STE from crowdsourced triplets (based on how the products work)

ing triplet embedding methods where not all of the triplet constraints being satisfied [36, 48]. This can be attributed to inconsistency across the crowd among other reasons. Using distance error, as shown in Figure 5b, GNMDS, STE, and t-STE demonstrate a level of convergence in embeddings at about 50 percent of collected triplets. At this point, the similarity scores change only slightly in comparison to the scores from the final embedding. CKL does not demonstrate convergence in either case and therefore, is not considered further.

Finally, the median rank correlation coefficient (Kendall's τ) of the product rankings is calculated across the methods, as shown in Figure 5c, to determine if there are differences in the rankings (i.e. relative similarities calculated from the 2D space) when using a specific triplet embedding method. There is a strong correlation between rankings across methods, demonstrating that in addition to performing similarly in terms of errors in satisfying triplet constraints, the different methods only have a small effect on the resulting pairwise similarities. A closer look at the 2D embedding in Figure 4 verifies that products that were expected to be close to each other in the 2D embedding space (e.g. the two vise grips, located in the upper middle area) are actually close to each other using t-STE.



(a) Average triplet generalization error across ten folds

(b) Mean squared error of distance matrix compared to distance matrix of final embedding

(c) Median rank correlation coefficient between rankings derived from comparable methods

Fig. 5: Triplet generalization error and distance error verify that the t-STE embedding is not significantly changing with the addition of new triplets after a point. The embedding technique does not significantly impact similarity values derived from the triplet embeddings.

4.2 Comparing human perception of functional similarity at different levels of abstraction

To probe if the participants' conception of functional similarity (for this set of products) is affected by changing abstraction level, defined here as the product's purpose vs. working mechanism, the learned embeddings for the two levels of abstraction are compared. The rank correlation of the upper triangular portion of the pairwise similarity matrix from the human-determined embeddings (considering two dimensionality hyperparameters), is shown in Table 2, displaying only a relatively high correlation across the two levels of abstraction (for τ = 0.54, there is a positive correlation between 77 percent of all possible pairs of ranks being compared). This reflects that the abstraction seems to affect human consideration of functional similarity *sometimes*. Rankings of each product's similarity relative to all other products are found to determine which products might be driving any differences. These rank correlations for each product across the levels of abstraction are also displayed in Table 2. The correlations range from weaker (e.g. the water pump) to stronger (e.g. the two types of vise grips) in comparison to the full embeddings, demonstrating that for some of the products, the rankings match regardless of how function is presented, while for others, the rankings differ significantly. Therefore, a significant divergence in how humans consider similarity when framing function around a product's purpose vs. working mechanism exists, but only for a smaller subset of products.

Table 2: Rank correlation coefficient (Kendall's τ) for full embeddings and each product when comparing human-evaluated similarities across functional abstraction levels. Shaded rows indicate the subset of products with a rank correlation coefficient below the median (0.585).

	Rank Corr. Coeff.	p-value
Full embeddings		
2 dimensions	0.54	<0.01
13 dimensions	0.38	<0.01
Product		
Toy Plane	0.31	0.07
Alcohol Detector	0.64	<0.01
All-in-one Printer	0.50	<0.01
Bike	0.65	<0.01
Blower/Vacuum	0.22	0.21
CD Player	0.68	<0.01
Drink Cooler	0.60	<0.01
DVD Player	0.72	<0.01
Nerf Gun	0.40	0.02
Game Controller	0.36	0.03
Power Razor	0.56	<0.01
Stapler	0.58	<0.01
Hulk Hands	0.33	0.05
Lawn Mower	0.42	0.01
Quick Grip Vise	0.73	<0.01
Scooter	0.65	<0.01
VHS Player	0.67	<0.01
Vise Grip	0.78	<0.01
Water Pump	0.16	0.37
Zip Disk Drive	0.59	<0.01

Since the participants were presented with the same 20 triplets (in a randomized order) for each level of abstraction, each participant's ratings are compared across levels to see if the triplets were answered in the same way. Participants answer a mean of 69 percent (SD: 14%, min: 35%, max: 95%) of the triplets in the same way across both conditions, again supporting that participants only sometimes answer differently when presented with the two types of function information and that it highly depends on the type of product.

4.3 Human judgements vs. quantitative measures using triplets and embeddings

Comparisons between human similarity judgements of product function and functional similarity computed through similarity measures are made using both the learned embeddings and triplets.

4.3.1 Matching collected triplets with generated triplets across abstraction level

The percent of human triplets that match with generated triplets at the lower abstraction is 58 percent as an average across all measures, indicating that the measures and human judgements agree for approximately 12 of the 20 triplets a participant provides. Matching between human and generated triplets at the higher abstraction level averages 55 percent. Matching between human and computed triplets, across all measures, is statistically significantly different for the lower abstraction and the higher abstraction according to a dependent t-test (t(5)=3.45, p=0.02). The result indicates that measures (at least those considered here) generally have improved agreement with human judgements at the lower abstraction level than at the higher abstraction level. Comparing individual measures, at the lower abstraction level (working mechanism), a repeated measures ANOVA does not show a statistically significant effect of the measure type on the percentage of matching triplets. However, at the higher abstraction level (purpose), a repeated measures ANOVA does show a statistically significant effect between measures(F(5, 335)=6.26, $p<<0.01$). A post-hoc analysis with a Tukey HSD correction at this abstraction level shows a significant effect in comparisons involving the spectral measure — the spectral measure with SMC (p=0.01), spectral measure with Jaccard similarity (p=0.03), and spectral measure with cosine similarity (p=0.02). The spectral measure therefore demonstrates higher alignment with the collected triplet ratings than any of the matching-based similarity measures when considering the higher level of abstraction, matching 59 percent of triplets from participants on average (on par with matching at a lower abstraction generally). In summary, none of the measures match human ratings very well, but they match human ratings at a lower abstraction better than at the higher abstraction level. One exception to this is when using the spectral measure, which compares functional models based on their topological similarity regardless of node labels (i.e. the specific sub-functions). The results support that network-based measures may be more useful for capturing what people consider to be functional similarity at a higher abstraction level.

4.3.2 Correlations between pairwise similarity matrices of human-determined embeddings and computed measures across abstraction level

The rank correlation of the upper triangular portion of each human-determined embedding's pairwise matrix and each measure's pairwise matrix is shown in Table 3, demonstrating an overall comparison of human and computational representations of functional similarity. Interpretation of the correlation values themselves indicates that at maximum (τ=0.30) and minimum (τ=0.05), there is a positive correlation for 65 and 53 percent, respectively, of all ranks pairs being compared between the results from a similarity measure vs. a human embedding. In general, the correlation values are higher for the lower abstraction level than for the higher abstraction level, again supporting that the measures are better aligned with humans at the lower abstraction than at the higher abstraction.

Looking more closely at the specific measures that align better, at the lower abstraction, the top two measures with the highest correlation are the NetSimile and Jaccard measures. At the higher abstraction, the two measures that have the highest correlation regardless of the embedding dimensionality are the spectral and NetSimile measures. For the higher abstraction embedding in two dimensions, the measure with the highest correlation (the spectral measure) corresponds to the results found by just using the raw triplets, indicating that perhaps the two dimensional embedding is sufficient to capture the human signal.

4.4 Product-level comparison considering similar vs. "highly similar"

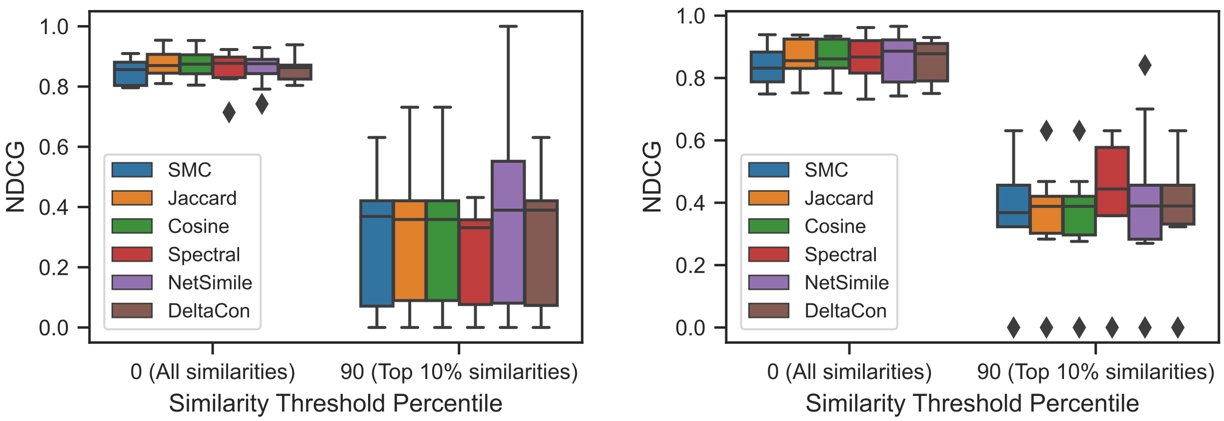
This section examines more closely how quantitative measures compare to the human-determined similarity when considering highly similar products vs. the entire product space. In this case, only

Table 3: Rank correlation of similarity matrices from human embedding with computed matrices. The shaded cells indicate the measure that has the highest correlation with the human embedding at the given abstraction and dimensionality. * and ** indicate significance at the 0.05 and 0.01 thresholds respectively.

	Human Embedding			
Abstraction	Lower		Higher	
Dimensionality	2	13	2	13
SMC	0.12**	0.18**	0.10*	0.08
Jaccard	0.23**	0.27**	0.10*	0.07
Cosine	0.20**	0.26**	0.09	0.05
Spectral	0.20**	0.23**	**0.21****	0.19**
NetSimile	**0.26****	**0.30****	0.18**	**0.20****
DeltaCon	0.11**	0.17**	0.11*	0.11*

the subset of products where the functional abstraction appears to affect functional similarity rating (a low correlation between rankings across abstraction as shown by the shaded rows in Table 2) is considered. As described in Section 3, each product has a ranking for how similar other products are to it. When thresholds are applied, the NDCG measure prioritizes alignment between humans and measures along the highly similar products (those passing the threshold). It should be noted that the threshold is applied across all product comparisons, but the NDCG is calculated independently for each product (a row or column in the similarity matrix). This means that a product can have any number of products that are functionally "similar enough" to it (including zero, as may be the case where the similarity threshold is very high)

Figure 6 shows the NDCG of each measure given reference rankings from the two-dimensional human-sourced embedding at each abstraction level. The NDCG measure reveals the similar information to both the triplet matching and the embedding correlations when there are no thresholds to similarity applied — at the lower abstraction, the measures appear to perform relatively *similarly*, while at the higher abstraction the spectral and NetSimile measures have better alignment with human ratings. When no thresholds are applied, the measures must perform well across all levels of similarity (not just highly similar cases). However, using the NDCG allows us to see the qualitative changes in which measure performs the best when only looking at highly functionally similar products. For example, at the lower abstraction level, considering only the most similar products, alignment between measures and human-determined rankings becomes highly dependent on the product, as demonstrated by the larger ranges of NDCG values. Then, at the higher level of abstraction, the median of the spectral measure is notably higher than the other measures when the threshold is at its highest, indicating that it aligns more closely with the human judgements regarding the highly functionally similar products. While these more fine-grained investigations cannot lead to definite conclusions, together with the other results they provide a clear indication of how similarity measures compare to human evaluations. The similarity measures align *similarly* with human perception of functional similarity (but not perfectly) unless it is important to find highly functionally similar products or operate at a specific level of abstraction. In those cases, differences among the measures arise, and some align with human representations better than others, often also depending on the specific product type.



(a) Lower abstraction (product working mechanism)

(b) Higher abstraction (product purpose)

Fig. 6: NDCG (comparing product-level similarity rankings from measures to “reference” product-level human similarity rankings) across levels of abstraction with and without a similarity threshold applied. The overall NDCG drops when a similarity thresholds are applied because fewer products have relevance greater than 0. Notably, the addition of thresholds reveals the product-level differences across individual measures in alignment with human rankings.

5 DISCUSSION

During design, it may be necessary to specify similarity along several dimensions to help facilitate connections across domains and move beyond surface-level similarities. However, it is difficult to incorporate this type of flexibility into measures of similarity that might be useful for design automation or support. Furthermore, to apply computational methods to design, it is important to understand when human decisions might be in conflict (or alignment) with computational support tools. In this work, these questions are explored in the context of functional similarity, a dimension of similarity particularly important for design, by directly comparing results from human judgements to those calculated from functional models. In addition, factors that lead to these conflicts, such as a threshold for similarity or different levels of abstraction, are studied. There are several key findings. First, varying abstraction level affects what people consider to be similar product functions to an extent. The quantitative measures considered here have a limited ability to capture the human representation of function, especially at the higher level of abstraction. We also find that no similarity measure consistently matches best with human ratings across both abstraction levels. Finally, differences between the individual measures' ability to align with human perception of functional similarity appear to depend on whether it is desirable that they align on what is highly similar vs. progressively less similar products. This is pertinent for analogical design, where it may be desirable to search beyond the space of highly similar products. These results are expanded upon further below.

5.1 Computed similarity may fail to capture human representations of functional similarity

In this work, crowdsourcing and triplet embedding is used to quantify how people consider products to be functionally similar. The human similarity judgement embedding is created as an aggregate across the participant population though in reality, individuals may perceive similarity in different ways, even when instructions specify consideration along a certain dimension. While many of the limitations to using embeddings mentioned by Ahmed et al. [36] still apply, the method provides a way to compare these judgements with what can be directly computed from functional representations of products without using a specific design task. The correlations between results from any of the six computed similarity measures and the human embeddings seem to indicate that the measures are not capturing human perception of functional similarity. This discrepancy is more pronounced when humans consider function at a higher abstraction, for almost all of the measures. Related work comparing human embeddings with the low-dimensional embeddings from image classification models (deep learning models as opposed to the quantitative measures here) also finds correlations on the order of those found here (the highest correlation is 0.30) [32]. This raises questions about whether the representations of function that these similarity measures can capture are *sufficient* for design applications, even if they do not align strictly with human representations. For instance, prior work indicates that people may use a structural alignment approach in similarity, and specifically notes that people tend to match common features across items, an attribute that is shared by the Jaccard and cosine similarity measure [34]. However, when considering functional similarity, unlike the measures applied to functional models, humans cannot easily match the numerous sub-functions, instead making a more holistic assessment. Therefore, at the lower abstraction, using results from the considered similarity measures may provide them with information they have missed. On the other hand, measures must be used carefully if applied when humans are operating at the higher abstraction level, as only the network-based measures demonstrate even a small amount of alignment with humans, which may not be enough.

5.2 Humans conceptualize functional similarity at different levels of abstraction and similarity measures may have limited ability to reflect this difference

The results indicate that the abstraction level can affect which products humans consider to be most functionally similar. Although it was expected that the embeddings would look significantly different for almost all products when considering the different levels of abstraction, it turns out that a smaller subset of products may drive the differences. An overlay of maps of the subset of products with low rank correlations (below the median) is shown in Figure 7.

From this map, a specific example of the effect of abstraction is in the trio of products including the Hulk Hands, Toy Plane, and Nerf Gun. The Hulk Hands product and the Toy Plane product are closer in the higher abstraction function map (their functions are described as providing sound and motion for entertainment respectively), while the Toy Plane product moves away from the Hulk Hands product and closer to the Nerf Gun product in the lower abstraction function map. This can

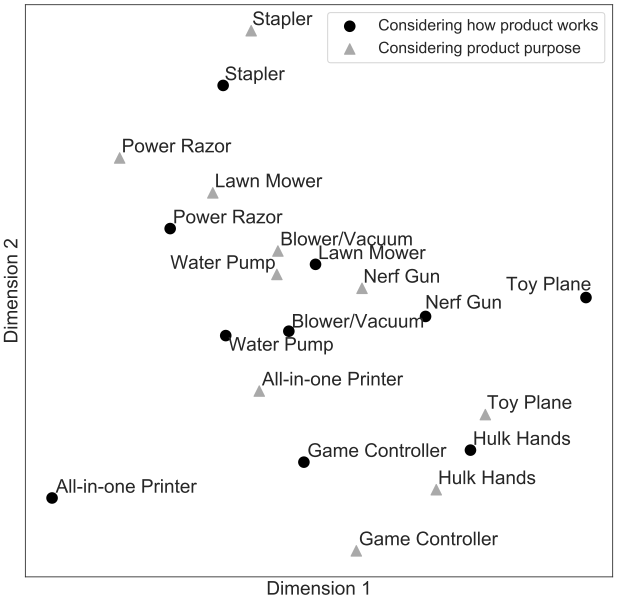


Fig. 7: Two-dimensional embeddings for the subset of products with low rank correlation coefficients across abstraction levels (visualized using Procrustes analysis in SciPy). Some products are brought closer together or pushed farther apart from each other depending on the level of functional abstraction considered.

potentially be explained by the shared pneumatic mechanism between the Toy Plane and Nerf Gun that is not considered for its overall purpose. It is noted that the function information presented to participants for the lower level of abstraction was summarized from the full functional model and therefore, not necessarily complete. By investigating a larger variety of products, it may be possible to understand the types of products for which abstraction level affects consideration of functional similarity and why. In addition, Chaudhari et al. [55] point out that how people view similarity is dynamic. This is an important consideration when looking at levels of abstraction, where the level of expertise may play a role in the ability to draw more abstract functional connections.

When comparing the human judgements with quantitative measures directly computed on products and including the factor of abstraction, there is a discrepancy in access to information: humans were provided with the higher-level function, while the measures still operated on the full, lower-level functional model. This discrepancy can be addressed by using pruning rules on the functional models to remove unimportant information as done by Caldwell and Mocko [12]. However, it may also be desirable for a computed measure to be able to infer the higher abstraction from lower-level attributes rather have to directly provide both levels of abstraction. From this perspective, it is notable that the measures that align the best with human judgements at the higher abstraction, both when using the embeddings and when directly matching triplets, are network-based (particularly the spectral network similarity measure aligning significantly better with humans than the three other feature-matching-type measures: SMC, Jaccard, and cosine). Thus, in comparison to other types of measures, a network-based measure has the potential to allow access to the higher abstraction without the effort needed to directly learn the latent space with large amounts of data. This might encode the aspect of how humans consider relations and sets of relations within items when making comparisons as proposed by Gentner and Markman [29].

5.3 The ways similarity measures align with human judgements may differ when considering similar vs. "highly similar" products

It appears that the way people consider highly similar items cannot be captured in the same way as how people consider similarity more broadly. It is possible that the embedding does not accurately capture how people think of the more dissimilar products, as they are specifically asked to select the more similar product in the triplet. Additionally, there are limitations in the thresholding approach due to the small number of products considered, meaning that certain products may not have had items within the dataset that were similar at all. Further investigation into similarity thresholds may correspond to finding products that are "far" but not "too far" in terms of analogical distance. When deciding to utilize a measure to search for far-field sources of inspiration, it is desirable to choose a measure that does not have the highest alignment with human similarity judgements to provide unexpected results, as indicated by Fu et al. [18]. However, if it is important for the similarity measures to return functionally similar products in the exact way that humans are considering functional similarity, the results indicate that none of the measures considered here can be recommended. At the same time, since humans can adapt their notion of similarity to search for connections between products, their conception of function and functional similarity might change if they are presented with the output from the similarity measures instead of asked to make their judgements independently. Therefore, the context dependence of human similarity judgements could possibly nudge alignment between humans and similarity measures closer than what is found in this study.

5.4 Implications for design

Research finds cognitive processes such as long-term memory retrieval, semantic/associative processes, and visual perception, to be relevant during design from both a design-as-search and a design-as-exploration perspective. In design practice, limitations of these cognitive processes might be addressed through automation [56]. It is possible, for instance, that the ability of humans to view similarity dynamically and measures to compute similarity in a structured way can complement each other. If humans assess similarity based on products that they have encountered before instead of only the set of products they were presented, expertise may affect what humans considered functionally similar. Those with more experience or higher expertise may have encountered more products to which they can compare. While the measures considered here can only compare relative to a set of products they have access to, this set may also be larger than what a person can remember. Additionally, in cases where functional connections between products are not very clear (perhaps the products are quite dissimilar), humans may resort to other dimensions to make their decisions, while computational measures can continue to make these judgements without this limitation. In such a case, perhaps humans should not be used as the standard to which functional similarity measures are held and the divergence can be exploited. Results from recent work on ratings of design concept novelty have also hinted at this point, finding subsets of highly-rated designs to differ across comparing human and computational evaluations [57]. Thus, future work might investigate ways to determine the effectiveness of combining human evaluations with quantitative evaluations in design more broadly, even beyond functional similarity.

While there are several ways to capture product function (e.g. functional models, descriptions, patents) in engineering design, leveraging this knowledge for creative transfer of ideas across domains will require methods to assist designers in searching through a design space along the dimension of function [6, 33]. For these methods to be adopted in design practice, functional similarity must be defined in a way that proves useful to humans. However, the findings in this study provokes consideration of whether this necessarily means computational and human similarity should be aligned. In this study, no considered measure demonstrated very close alignment with human perception of functional similarity. Only one measure (NetSimile) demonstrated some robustness to considerations like abstraction. If computational output must be highly in line with humans' representations, more adaptive ways of determining functional similarity are required. On the other hand, further investigation could reveal measures that transcend the limitations and instead augment human similarity perception. Then, if a designer could flexibly retrieve products or ideas that were functionally similar at the right level of abstraction, and at the right "amount" of similar, perhaps they could make better use of the vast number of existing products to inspire new ideas.

6 CONCLUSION

In this paper, human similarity judgements of functional similarity, using a set of consumer products, were crowdsourced and a triplet embedding method was applied to quantify these human judgments in a low-dimensional embedding space. This representation provides insight into the alignment between how humans view functional similarity and how these functional similarities can be directly computed from the products. The results indicate that human and computational representations of functional similarity diverge and are affected by different ways that humans might consider similarity. The way highly similar products are considered by humans compared to "somewhat similar" products may not be captured by these existing measures, affecting applications such as design-by-analogy, where analogical distance must be controlled. Additionally, for some products, the level of abstraction can influence whether human judgements align with computational measures. Factoring in higher functional abstraction, network-based measures that account for relations between elements may be appropriate. These types of measures can potentially be used to represent how humans abstract function when it is not possible to directly learn a measure from a large quantity of data collected from humans. Further work is needed to better define functional similarity in a way that is interpretable and useful to humans across different abstraction levels.

ACKNOWLEDGEMENTS

This work has been supported by the Regents of the University of California. The findings presented in this work represent the views of the authors and not necessarily those of the sponsors.

We would like to thank the authors of [48] for making their code available. A preliminary version of this study was presented for ASME IDETC 2021 [58]. We thank the reviewers for their constructive comments.

REFERENCES

[1] Christensen, B. T., and Schunn, C. D., 2007. "The relationship of analogical distance to analogical function and preinventive structure: The case of engineering design". *Memory & Cognition,* **35**(1), Jan., pp. 29–38.

[2] Ulrich, K. T., and Eppinger, S. D., 2004. *Product Design and Development*. McGraw-Hill/Irwin.

[3] Goucher-Lambert, K., Moss, J., and Cagan, J., 2019. "A neuroimaging investigation of design ideation with and without inspirational stimuli—understanding the meaning of near and far stimuli". *Design Studies,* **60**, Jan., pp. 1–38.

[4] Linsey, J. S., Laux, J., Clauss, E. F., Wood, K. L., and Markman, A. B., 2007. "Effects of Analogous Product Representation on Design-By-Analogy". In DS 42: Proceedings of ICED 2007, the 16th International Conference on Engineering Design, Paris, France, 28.-31.07.2007, pp. 337–338 (exec. Summ.), full paper no. DS42_P_477.

[5] Linsey, J. S., Wood, K. L., and Markman, A. B., 2008/ed. "Modality and representation in analogy". *AI EDAM,* **22**(2), pp. 85–100.

[6] Fu, K., Murphy, J., Yang, M., Otto, K., Jensen, D., and Wood, K., 2015. "Design-by-analogy: Experimental evaluation of a functional analogy search methodology for concept generation improvement". *Research in Engineering Design,* **26**(1), Jan., pp. 77–95.

[7] McAdams, D. A., and Wood, K. L., 2002. "A Quantitative Similarity Metric for Design-by-Analogy". *Journal of Mechanical Design,* **124**(2), June, pp. 173–182.

[8] Turner, C., and Linsey, J., 2016. "Analogies from Function, Flow and Performance Metrics". In Workshop Proceedings from the 24th International Conference on Case Based Reasoning.

[9] Taylor, L. E., and Henderson, M. R., 1994. "The Roles of Features and Abstraction in Mechanical Design". In ASME 1994 Design Technical Conferences Collocated with the ASME 1994 International Computers in Engineering Conference and Exhibition and the ASME 1994 8th Annual Database Symposium, American Society of Mechanical Engineers Digital Collection,

pp. 131–140.

[10] Maier, J. F., Eckert, C. M., and Clarkson, P. J., 2017/ed. "Model granularity in engineering design – concepts and framework". *Design Science,* **3**.

[11] Caldwell, B. W., Thomas, J. E., Sen, C., Mocko, G. M., and Summers, J. D., 2012. "The Effects of Language and Pruning on Function Structure Interpretability". *Journal of Mechanical Design,* **134**(6), Apr.

[12] Caldwell, B. W., and Mocko, G. M., 2011. "Functional Similarity at Varying Levels of Abstraction". In ASME 2010 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, American Society of Mechanical Engineers Digital Collection, pp. 431–441.

[13] Kittur, A., Yu, L., Hope, T., Chan, J., Lifshitz-Assaf, H., Gilon, K., Ng, F., Kraut, R. E., and Shahaf, D., 2019. "Scaling up analogical innovation with crowds and AI". *Proceedings of the National Academy of Sciences,* **116**(6), Feb., pp. 1870–1877.

[14] Hope, T., Tamari, R., Kang, H., Hershcovich, D., Chan, J., Kittur, A., and Shahaf, D., 2021. "Scaling Creative Inspiration with Fine-Grained Functional Facets of Product Ideas". *arXiv:2102.09761 [cs]*, Feb.

[15] Anandan, S., Teegavarapu, S., and Summers, J. D., 2006. "Issues of Similarity in Engineering Design". In ASME 2006 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, American Society of Mechanical Engineers Digital Collection, pp. 73–82.

[16] Ranawat, A., and Hölttä-Otto, K., 2009. "Four dimensions of design similarity". In ASME 2009 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, IDETC/CIE2009, pp. 1069–1077.

[17] Fischer, M. S., Holder, D., and Maier, T., 2020. "Evaluating Similarities in Visual Product Appearance for Brand Affiliation". In Advances in Affective and Pleasurable Design, S. Fukuda, ed., Advances in Intelligent Systems and Computing, Springer International Publishing, pp. 3–12.

[18] Fu, K., Chan, J., Schunn, C., Cagan, J., and Kotovsky, K., 2013. "Expert representation

of design repository space: A comparison to and validation of algorithmic output". *Design Studies,* **34**(6), Nov., pp. 729–762.

[19] Kwon, E., Huang, F., and Goucher-Lambert, K., 2021. "Multi-modal search for inspirational examples in design". In International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, Vol. 85420, American Society of Mechanical Engineers, p. V006T06A020.

[20] Stone, R. B., and Wood, K. L., 2000. "Development of a Functional Basis for Design". *Journal of Mechanical Design,* **122**(4), Dec., pp. 359–370.

[21] Hirtz, J., Stone, R., Mcadams, D., Szykman, S., and Wood, K., 2002. "A Functional Basis for Engineering Design: Reconciling and Evolving Previous Efforts". *Research in Engineering Design,* **13**, Mar., pp. 65–82.

[22] Agyemang, M., Linsey, J., and Turner, C. J., 2017. "Transforming functional models to critical chain models via expert knowledge and automatic parsing rules for design analogy identification". *AI EDAM,* **31**(4), Nov., pp. 501–511.

[23] Ferrero, V. J., Alqseer, N., Tensa, M., and DuPont, B., 2020. "Using Decision Trees Supported by Data Mining to Improve Function-Based Design". In ASME 2020 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, American Society of Mechanical Engineers Digital Collection.

[24] Mikes, A., Edmonds, K., Stone, R., and DuPont, B., 2020. "Optimizing an Algorithm for Data Mining a Design Repository to Automate Functional Modeling". In ASME 2020 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference.

[25] Nandy, A., Dong, A., and Goucher-Lambert, K., 2021. "Evaluating Quantitative Measures for Assessing Functional Similarity in Engineering Design". *Journal of Mechanical Design,* **144**(3), Sept.

[26] Hebart, M. N., Zheng, C. Y., Pereira, F., and Baker, C. I., 2020. "Revealing the multidimensional mental representations of natural objects underlying human similarity judgements". *Nature Human Behaviour,* **4**(11), Nov., pp. 1173–1185.

[27] Tversky, A., 1977. "Features of similarity". *Psychological Review,* **84**(4), pp. 327–352.

[28] Goldstone, R. L., Medin, D. L., and Halberstadt, J., 1997. "Similarity in context". *Memory & Cognition,* **25**(2), Mar., pp. 237–255.

[29] Gentner, D., and Markman, A. B., 1997. "Structure Mapping in Analogy and Similarity". *American Psychologist*, p. 12.

[30] Towne, W. B., Rosé, C. P., and Herbsleb, J. D., 2016. "Measuring Similarity Similarly: LDA and Human Perception". *ACM Transactions on Intelligent Systems and Technology,* **8**(1), Oct., pp. 1–28.

[31] Ellis, D., Whitman, B., Berenzweig, A., and Lawrence, S., 2002. "The Quest for Ground Truth in Musical Artist Similarity.". In Proceedings of the 3rd International Conference on Music Information Retrieval.

[32] Roads, B. D., and Love, B. C., 2021. "Enriching ImageNet With Human Similarity Judgments and Psychological Embeddings". In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3547–3557.

[33] Gill, A. S., Tsoka, A. N., and Sen, C., 2019. "Dimensions of Product Similarity in Design by Analogy: An Exploratory Study". In ASME 2019 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, American Society of Mechanical Engineers Digital Collection.

[34] McTeague, C. P., Duffy, A., Hay, L., Vuletic, T., Campbell, G., Choo, P. L., and Grealy, M., 2015. "Insights into design concept similarity judgements". In 15th International Design Conference, pp. 2087–2098.

[35] Hwang, D., and Wood, K., 2020. "Assessing the Novelty of Design Outcomes: Using a Perceptual Kernel in a Crowd-sourced Setting". In Ninth International Conference on Design Computing and Cognition.

[36] Ahmed, F., Ramachandran, S. K., Fuge, M., Hunter, S., and Miller, S., 2018. "Interpreting Idea Maps: Pairwise Comparisons Reveal What Makes Ideas Novel". *Journal of Mechanical Design,* **141**(021102), Dec.

[37] Siangliulue, P., Arnold, K. C., Gajos, K. Z., and Dow, S. P., 2015. "Toward Collaborative

Ideation at Scale: Leveraging Ideas from Others to Generate More Creative and Diverse Ideas". In Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing, ACM, pp. 937–945.

[38] Vargas, S., and Castells, P., 2011. "Rank and relevance in novelty and diversity metrics for recommender systems". In Proceedings of the Fifth ACM Conference on Recommender Systems, RecSys '11, Association for Computing Machinery, pp. 109–116.

[39] Shah, J. J., Smith, S. M., and Vargas-Hernandez, N., 2003. "Metrics for measuring ideation effectiveness". *Design Studies,* **24**(2), Mar., pp. 111–134.

[40] Goucher-Lambert, K., and Cagan, J., 2019. "Crowdsourcing inspiration: Using crowd generated inspirational stimuli to support designer ideation". *Design Studies,* **61**, Mar., pp. 1–29.

[41] Srinivasan, V., Song, B., Luo, J., Subburaj, K., Elara, M. R., Blessing, L., and Wood, K., 2018. "Does Analogical Distance Affect Performance of Ideation?". *Journal of Mechanical Design,* **140**(071101), May.

[42] Chan, J., Dow, S. P., and Schunn, C. D., 2015. "Do the best design ideas (really) come from conceptually distant sources of inspiration?". *Design Studies,* **36**, Jan., pp. 31–58.

[43] Fu, K., Chan, J., Cagan, J., Kotovsky, K., Schunn, C., and Wood, K., 2013. "The Meaning of "Near" and "Far": The Impact of Structuring Design Databases and the Effect of Distance of Analogy on Design Output". *Journal of Mechanical Design,* **135**(2), Feb.

[44] Goel, V., 1995. "Cognitive processes involved in design problem solving". In *Sketches of Thought*. MIT Press, Oct., pp. 95–126.

[45] , 2020. The Design Repository. http://ftest.mime.oregonstate.edu/repo/browse/.

[46] Tamuz, O., Liu, C., Belongie, S., Shamir, O., and Kalai, A. T., 2011. "Adaptively learning the crowd kernel". In Proceedings of the 28th International Conference on International Conference on Machine Learning, ICML'11, Omnipress, pp. 673–680.

[47] Agarwal, S., Wills, J., Cayton, L., Lanckriet, G., Kriegman, D., and Belongie, S., 2007. "Generalized Non-metric Multidimensional Scaling". In Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics, PMLR, pp. 11–18.

[48] van der Maaten, L., and Weinberger, K., 2012. "Stochastic triplet embedding". In 2012 IEEE

International Workshop on Machine Learning for Signal Processing, IEEE, pp. 1–6.

[49] Wills, P., and Meyer, F. G., 2020. “Metrics for graph comparison: A practitioner’s guide”. *PLOS ONE,* **15**(2), Feb., p. e0228728.

[50] Berlingerio, M., Koutra, D., Eliassi-Rad, T., and Faloutsos, C., 2013. “Network similarity via multiple social theories”. In Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, pp. 1439–1440.

[51] Koutra, D., Vogelstein, J. T., and Faloutsos, C., 2013. “Deltacon: A principled massive-graph similarity function”. In Proceedings of the 2013 SIAM International Conference on Data Mining, SIAM, pp. 162–170.

[52] Hagberg, A. A., Schult, D. A., and Swart, P. J., 2008. “Exploring Network Structure, Dynamics, and Function using NetworkX”. In Proceedings of the 7th Python in Science Conference, pp. 11–16.

[53] Kriegeskorte, N., Mur, M., and Bandettini, P., 2008. “Representational similarity analysis - connecting the branches of systems neuroscience”. *Frontiers in Systems Neuroscience,* **2**, p. 4.

[54] Wang, Y., Wang, L., Li, Y., He, D., Liu, T.-Y., and Chen, W., 2013. “A Theoretical Analysis of NDCG Type Ranking Measures”. In Proceedings of the 26th annual conference on learning theory (COLT 2013), Vol. 8, p. 6.

[55] Chaudhari, A. M., Bilionis, I., and Panchal, J. H., 2019. “Similarity in Engineering Design: A Knowledge-Based Approach”. In ASME 2019 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, American Society of Mechanical Engineers Digital Collection.

[56] Hay, L., Duffy, A. H. B., McTeague, C., Pidgeon, L. M., Vuletic, T., and Grealy, M., 2017. “A systematic review of protocol studies on conceptual design cognition: Design as search and exploration”. *Design Science,* **3**, p. e10.

[57] Camburn, B., He, Y., Raviselvam, S., Luo, J., and Wood, K., 2020. “Machine Learning-Based Design Concept Evaluation”. *Journal of Mechanical Design,* **142**(3), Jan.

[58] Nandy, A., and Goucher-Lambert, K., 2021. “Aligning Human and Computational Evaluations

of Functional Design Similarity". In ASME 2021 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, American Society of Mechanical Engineers Digital Collection.

APPENDIX A PRODUCTS AND FUNCTION DESCRIPTIONS

Product	Product purpose (higher abstraction)	How product works (lower abstraction)
Toy Plane	Provide motion for entertainment	Humans pump pressurized air into the plane and throw it to give the plane translational motion. The propellors rotate.
Alcohol Detector	Measure alcohol	Humans turn on the device and blow into it. The device collects the breath sample and uses a chemical reaction to determine and display the alcohol level.
All-in-one Printer	Transform paper	Humans turn the printer on and insert paper. Print data is imported to the printer and then electrical energy is used to signal the printer to release the stored liquid ink. The ink changes the blank paper to the printed paper and the print status is displayed. A scanned document is converted to a signal and exported as scan data.
Bike	Transfer human	Humans pedal to provide mechanical energy for translational motion to transport themselves.
Blower/ Vacuum	Import air and debris and expel air	Humans turn on the device and electrical energy is used to signal the blower or vacuum setting. Air is expelled in the blower setting. An air and debris mixture is taken in and the debris are stored in the vacuum setting.
CD Player	Read a CD	Humans insert a CD and turn on the player. Electrical energy is used to start mechanical rotation of the CD and the lens focuses electromagnetic energy (laser) on the moving disk to read it and play the relevant audio. Buttons are used to control other actions such as pause and repeat.
Drink Cooler	Transfer thermal energy	Humans place the device on a surface and places a cup on top. Electrical energy is used to start mechanical rotation of a fan and extract heat. The fan expels air and the heat is transferred out.
DVD Player	Read a DVD	Humans insert a DVD and turn on the player. Electrical energy is used to start mechanical rotation and the lens focuses electromagnetic energy (laser) on the moving disk to read it. The electromagnetic energy is changed to electrical energy, which is used to display the video and play audio. Buttons are used to control other actions such as pause and eject.
Nerf Gun	Export ammo	Humans load the ammo, pump air into the gun, and pull a mechanical trigger. The pressurized air causes translational motion of the ammo and the gun emits noise.
Game Controller	Control computer	Humans push mechanical buttons or directional joysticks to actuate an electric signal. The electric signal is turned into a control signal that sends data to the connected electronic device as well as into electromagnetic energy (light) and mechanical vibration on the controller.

Power Razor	Separate hair from human	Humans provide translational motion to the razor over the surface of their skin through their hands. Electrical energy is converted to mechanical energy in the razor to cut the hair and separate it from the surface of the skin. The razor releases the cut hair, heat, and noise.
Stapler	Couple paper	Humans store staples in the stapler. Paper is positioned between the top and bottom housing of the stapler and force is applied to the top housing by the hand. The staple is separated from other staples and couples the sheets of paper together. The stapler releases the stapled pages and noise.
Hulk Hands	Emit sound for entertainment	Humans place their hands in the gloves. The gloves detect and process an electrical signal from human movement. The electrical signal is converted to noise.
Lawn Mower	Separate grass from ground	Humans push the lawn mower to add translational motion and turn it on. Liquid fuel is stored and the chemical energy in it is converted to mechanical energy. The mechanical energy is used to cut the grass and expel the cut grass pieces. The lawn mower releases heat, noise, and fumes.
Quick Grip Vise	Secure solid	Humans position the object and secure it by applying force to clamp it.
Scooter	Transfer human	Humans provide or stop translational motion to transport themselves.
VHS Player	Read a VHS tape	Humans turn on the player insert the tape, which is sensed and then guided in. Electrical energy is used mechanically translate the tape and then to start mechanical rotation of the wheels. The magnetic tape reel is read and encoded into video and audio signals, which are played. Electrical energy is also converted to electromagnetic energy (light) to indicate the status. Buttons are used to control other signals such as stop and eject.
Vise Grip	Secure solid	Humans position the object and secure it by applying force, changing its status from unclamped to clamped.
Water Pump	Move liquid	Humans turn the pump on. Electrical energy is converted to mechanical energy and then to pressurized air within the pump, which moves the liquid. Heat, noise, and pressurized air are released.
Zip Disk Drive	Read a zip disk	Humans turn on the reader and insert a zip disk, which is sensed and guided in. Electrical energy is converted to mechanical energy to rotate the disk and to actuate translation for the reading head. The magnetic energy from the disk is converted to electrical energy and is exported as data.

LIST OF TABLES

LIST OF FIGURES