INTERNATIONAL DESIGN CONFERENCE – DESIGN 2024
https://doi.org/10.1017/pds.2024.194

Datasets in design research: needs and challenges and the role of AI and GPT in filling the gaps

Mohammad Arjomandi Rad ✉, Tina Hajali, Julian Martinsson Bonde, Massimo Panarotto, Kristina Wärmefjord, Johan Malmqvist and Ola Isaksson

Chalmers University of Technology, Sweden

✉ radmo@chalmers.se

Abstract

Despite the recognized importance of datasets in data-driven design approaches, their extensive study remains limited. We review the current landscape of design datasets and highlight the ongoing need for larger and more comprehensive datasets. Three categories of challenges in dataset development are identified. Analyses show critical dataset gaps in design process where future studies can be directed. Synthetic and end-to-end datasets are suggested as two less explored avenues. The recent application of Generative Pretrained Transformers (GPT) shows their potential in addressing these needs.

Keywords: data-driven design, dataset gap, artificial intelligence (AI), design research, generative design

1. Introduction

As the synergy between Artificial Intelligence (AI) and Design Science (DS) accelerates (Moghaddam et al., 2023), the demand for large, well-annotated datasets has become increasingly critical. This growing need echoes the past importance of design catalogues (like the ones developed by German researchers in the 50s to 70s), which were collections of design solutions, principles, and patterns that have been identified and organized based on certain criteria. (Birkhofer et al., 2012). Datasets in DS, in their modern form, are defined as structured collections of data (Renear et al., 2010) specifically tailored for testing new processes and validating methodologies and processes. However, designers often struggle to find extensive suitable public datasets. There need to create open-source, high-quality datasets that are reflective of the diverse challenges in design. This fact has been emphasized by the design community (Regenwetter et al., 2022; Siddharth et al., 2022).

Today realistic, varied, and privacy-compliant datasets are not just a necessity but a strategic asset for idea or concept evaluation. The range of answers to such need is wide from some researchers calling for collaboration between design and data scientists to solve the issue (Chiarello et al., 2021), to others that stress the significance of maintaining lean and small datasets while enhancing quality features through feature engineering methodologies (Rad et al., 2022).

This paper aims to answer the question regarding the availability of open-source, high-quality datasets in design research. By design research we limit ourselves to common design process steps in product development model (Ulrich et al., 2008). Motivation for this research also comes from design activities in technologically advanced and security-intensive industries such as space (Panarotto et al., 2022) that are challenged by a lack of dataset. Employing recent advances in large language models like Generative Pretrained Transformers (GPT) for the creation of design datasets can potentially open a new era for

more efficiency and innovation. However, to harness the full potential of AI in creating datasets, a set of best practices must be established (Picard et al., 2023).
In this paper, first we review and categorize the landscape of datasets in design research, to explore the existing datasets in design research and highlight the gap. Next, we review what the barriers are in creating high-quality datasets in the design field. Further, we explore potential solutions and highlight the role of AI and GPT in addressing the gap. The paper is later concluded with future research recommendations.

2. Background

2.1. The existing datasets in design research

The realm of product design research is enriched by a diverse spectrum of datasets. Datasets found in this research were the result of employing targeted searches in Scopus and Google Scholar and analysis of selected sources in the research context. We have categorized 25 of these datasets, shown in Table 1, into six primary groups, based on the varied stages of product development as well as the the purpose each one is serving. The presented categories play a vital role in the complete understanding and application of design principles in research and real-world scenarios.

Table 1. Diverse dataset landscape in design research

Group / Focus	Name	Type	Reference
G1 Focuses on initial conceptualization and ideation	OpenSketch	Design Sketches	(Gryaditskaya et al., 2019)
	Quick, Draw!	Hand drawn sketch	(Xu et al., 2018)
	CADSketchNet	Sketch to CAD	(Manda et al., 2021)
	Sketch2Mesh	2D sketches to 3D shape	(Guillard et al., 2021)
	SPARE3D	3-view line drawings	(Han et al., 2020)
	DeepPatent2	Tech drawing from patents	(Ajayi et al., 2023)
G2 Focuses on generic 3D models and objects, provided in broad categories	ModelNet	3D CAD models	(Wu et al., 2015)
	PartNet	3D objects	(Mo et al., 2019)
	ShapeNet	3D CAD models	(Chang et al., 2015)
	ABC	3D CAD models	(Koch et al., 2019)
	MCB	3D objects	(Kim et al., 2020)
	ESB	3D CAD models	(Jayanti et al., 2006)
	Fusion 360	3D CAD model	(Willis et al., 2021)
G3 Provide many variants of a geometry for specific products	Airfoil coordinates	2D CAD NURBS	(Chen et al., 2019)
	Honda automobile hood	3D CAD models	(Ramnath et al., 2019)
	Airbag CAD design	2D CAD and screenshots	(Arjomandi Rad et al., 2023)
	SHIP-D	CAD models	(Bagazinski and Ahmed, 2023)
	BIKED: Computational bicycle design	Assembly and component images, design parameters	(Regenwetter et al., 2021)
G4 Topology optimization on a design solution	SimJEB: Jet Engine Bracket	3D Topology Optimization	(Whalen et al., 2021)
	Moving Morphable Bars (MBB)	2D Geometry-based Topology Optimization	(Hoang et al., 2022)
G5 Focuses on materials of different sorts	Mechanical MNIST	Heterogeneous materials	(Lejeune, 2020)
	METASET	3D unit cells	(Chan et al., 2021)
	Truss metamaterials	Metamaterials in the form of truss lattices	(Zheng et al., 2023)
G6 Other datasets	LINKS	Planar Linkage Mechanisms	(Heyrani Nobari et al., 2022)
	Brushstroke	Images of Artistic Strokes	(Shugrina et al., 2022)

https://doi.org/10.1017/pds.2024.194 Published online by Cambridge University Press

Despite the rich variety of datasets available in design research over the years, a significant gap persists, particularly in the early phases of the design process. As can be inferred from the table, current datasets predominantly focus on later stages of design development, such as detailed sketches, CAD models, and material properties. However, there is a notable gap in the collection and analysis of initial design stages, such as the compilation of user requirements, their corresponding design solutions (means), and the mapping of user needs to functional requirements as well as to those means. Addressing this gap by developing datasets that document and analyze the early stages of design, including requirement gathering and initial concept development, is crucial for enhancing the effectiveness and user-centeredness of design solutions. Such datasets would not only fill a crucial gap in design research but also provide a more holistic view of the product development process.

2.2. The need for more datasets

In the realm of data-driven design, the adequacy and depth of datasets play a pivotal role in advancing methodologies and achieving practical outcomes. However, a review of recent literature reveals a consensus among experts that the current state of datasets in this field is lacking in several key aspects. After gathering 12 review papers on the design and data set topic it turned out that 8 papers directly mention the need for more datasets. Table 2 compiles eight insightful quotes from each of these review papers in the data-driven design field. They highlight the need for datasets that are not only larger and more diverse but also of higher quality and completeness. Comparisons with other fields like computer vision reveal a stark contrast in the availability of well-annotated, public datasets. Furthermore, the integration of machine learning and NLP in design research is hampered by the lack of suitable datasets, particularly those that can serve as a standard for algorithm training and evaluation.

Table 2. Review papers from design science echoing the need for design datasets

Quote	Reference
"To shorten long training times, complete and noise-free design datasets created under suitable conditions are required."	(Yüksel et al., 2023)
"Most of the research … seldom has clarified how to prepare the specific dataset and how to conduct design knowledge-related feature engineering to identify the key design features that are supposed to be learned by the algorithm models."	(Yang et al., 2023)
"A Need for Large Multi-modal Design Datasets". "The community should collaboratively construct and maintain expansive design datasets with high-quality labels."	(Song et al., 2023)
"ML-based models can significantly aid in acquiring the massive number of datasets required for typical uncertainty quantification (UQ) procedures, which might not be practical to obtain from simulations and experiments."	(Babu et al., 2023)
"We propose that scholars develop standard datasets using design text as a common evaluation platform for future NLP applications.", "None of the NLP contributions that we have reviewed in this article leverage a design-specific gold standard dataset for evaluation."	(Siddharth et al., 2022)
"Compared to other research fields like Computer Vision, which have massive publicly available datasets, the availability of large, well-annotated, public datasets in engineering is severely lacking."	(Regenwetter et al., 2022)
"If the ML-enabled design automation is to be attained, larger datasets of real-world designs should be made freely available. Most of the ML algorithms reviewed herein have used training datasets in the order of the hundreds."	(Málaga-Chuquitaype, 2022)
"To foster the use of data in the context of engineering design, scholars and practitioners may develop packages especially designed for the ED context, as well as examples and datasets associated to particular methods."	(Chiarello et al., 2021)

The perception of dataset adequacy in the design field has faced a shift. Previously the design engineers tried to avoid large datasets as they were perceived less manageable. The number of sampling points was advised to be limited to $(n+1)(n+2)/2$, for n design variables (Koch et al., 1999; Wang and Shan, 2006). More than two decades after the early data-driven design approaches, scarcity of data seems to be now the limiting factor as the full-scale models require more than hundreds and thousands of sample

https://doi.org/10.1017/pds.2024.194 Published online by Cambridge University Press

points (Málaga-Chuquitaype, 2022). This mirrors the paradigm shift in overall Artificial Intelligence (AI) literature and reflects the recent success of generative models that utilize more complex algorithms that utilize much bigger datasets.

3. The barriers to creating datasets

After acknowledging the need for better datasets, here we talk about why there is a lack of datasets in design research. Acquiring datasets in product development comes with many complications. By reviewing the selected literature that presents a dataset or talks about the need for it, we found many challenges that can be categorized for better understanding. Figure 1 shows these categories in detail.

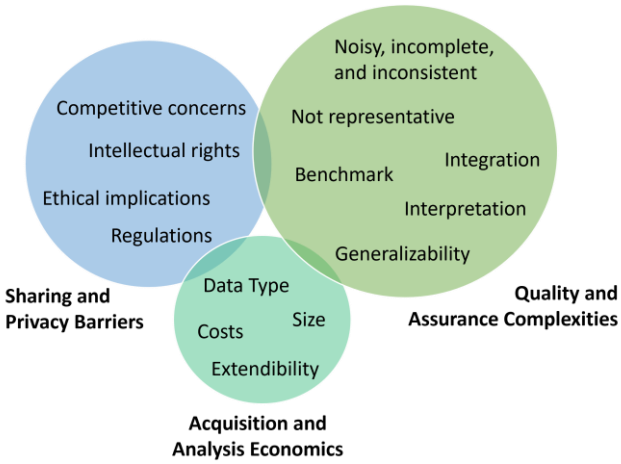

Figure 1. Three categories of identified barriers in creating design datasets

3.1. Sharing and privacy barriers

Often, the most relevant data is held by companies or institutions that are reluctant to share due to competitive concerns (Picard et al., 2023). This restricts access to valuable data that could inform product design and concept evaluations. Moreover, each industry has its own set of regulations regarding the use of data. Staying compliant with all applicable laws and regulations when collecting and using data is a constant challenge. On the other hand, There can be ethical implications for using real-world data, especially when it involves monitoring user behavior or collecting sensitive information (Goodman, 2014). Ensuring that the use of such data aligns with ethical standards is crucial. With stringent regulations such as the European "General Data Protection Regulation (GDPR), obtaining real-world data that contains personal information can be legally and ethically challenging (Gorkovenko et al., 2020). Companies must ensure data anonymization and secure handling to protect user privacy, which can be a complex and resource-intensive process when it comes to designing specific data types.

3.2. Quality assurance complexities

Real-world data can be noisy, incomplete, and inconsistent, making it difficult to use for accurate evaluations (Philip Chen and Zhang, 2014). Ensuring the data is clean, relevant, and of high quality requires substantial preprocessing, which can be both time-consuming and technically challenging. Additionally, real-world data comes from a variety of sources in different formats, which can make integration a complex task. Disparate data systems and incompatible formats can lead to significant challenges in data harmonization (Panchal et al., 2019).

With access to real-world data, the interpretation of that data can be complex (Zhu and Luo, 2023). It requires expertise in data science as well as deep domain knowledge to draw accurate conclusions that can inform the design and development process for making it more interpretable. Nevertheless, there is always a risk that the available data is not representative of the entire design space or use case scenarios (Regenwetter et al., 2022), leading to biased conclusions and insights. Ensuring diversity and mitigating bias in data sets is a key challenge. Furthermore, research faces big obstacles in creating diverse datasets across different design tasks. Simultaneously, the absence of established benchmarks in this exploratory

https://doi.org/10.1017/pds.2024.194 Published online by Cambridge University Press

field necessitates developing new metrics and incorporating broader experiments, possibly with human evaluations, to validate these models' effectiveness in real-world scenarios (Zhu et al., 2023; Zhu and Luo, 2023). For assessing dataset qualities, Pikard et al. suggest four core attributes: compatibility with existing data and models, completeness in capturing all necessary design characteristics, compactness to enhance processing efficiency, and generality to facilitate reuse across various applications, thus optimizing the cost of dataset creation (Picard et al., 2023).

3.3. Acquisition and analysis economics

The small size of existing datasets especially in Deep Generative Models (DGMs) is another challenge. It is well-known that recent advances in deep learning, especially in fields like computer vision and natural language processing, have been driven by billions of samples, a scale that is scarce in design research (Picard et al., 2023; Regenwetter et al., 2022). Acquiring or analyzing such a vast volume of high-quality, relevant data can also be expensive. The costs associated with data collection, storage, processing, and analysis can be prohibitive, especially for small and medium-sized enterprises.

Moreover, datasets suffer from extensibility which means they must be designed with the future in mind, allowing for updates and expansions as new data becomes available or as the design requirements evolve. This adaptability is essential for maintaining relevance and accuracy over time. Additionally, the variety in data types - from structured numerical data to unstructured text or images - requires versatile processing techniques, which further complicates the development and application of DGMs across the diverse landscape of design research (Zhu and Luo, 2023).

4. Future needs of datasets in design

In this section, we go through two types of datasets that are less discussed in and less known in the product development realm, based on the presented literature in previous sections.

4.1. Synthetic datasets

Synthetic, Mock, Pseudo, and Dummy datasets are all used in literature interchangeably. Synthetic dataset is a more common term and refers to algorithmically generated data that resemble real data closely, for example, well well-known MNIST dataset of handwritten digits is often used in its synthetic form for training machine learning models. Pseudo datasets are derived from or based on real data but have been modified or transformed. This transformation could be for reasons such as anonymization, simulation, or creating scenarios that are not present in the original dataset. Mock and Dummy are less common phrases and usually refer to custom generated for specific testing scenarios.

In the software development realm, libraries such as *Faker* and *Mockaroo* can fabricate realistic data across numerous categories including names, addresses, and credit card numbers (Curtis and Oliveira, 2023) in multiple formats such as CSV, JSON, SQL, and Excel (Ejlli, 2022). In system engineering, Test Data Management (TDM) systems are processes and tools used to manage and generate test data (Kasturi, 2020). These tools are particularly at populating complex datasets with representative data.

For product development, there is no clear differentiation for these datasets. Most of the datasets introduced in the early section can be considered synthetic as they are algorithmically created models of real products. Few efforts have also been published to create Pseudo datasets employing existing data augmentation methods (Du et al., 2021). These types of datasets can hold great potential for product design and development processes. By utilizing these datasets, product developers can potentially preserve the confidentiality of sensitive information. For instance, when working with proprietary designs or consumer data. Creating synthetic enables testing and refinement of products under diverse conditions at a considerably low cost, leading to improved design and functionality. These datasets can help in making informed decisions on user behavior, product performance under various conditions, and market trends, even when real data is scarce. By enabling rapid testing and iteration, synthetic datasets can significantly shorten the time of the task at hand, providing a competitive edge in fast-paced industries.

Synthetic datasets can be created to model consumer behavior and preferences. This can include virtual focus groups, simulated market responses to product changes, and predictive analytics for future trends.

https://doi.org/10.1017/pds.2024.194 Published online by Cambridge University Press

In industries like automotive, synthetic datasets can simulate how different materials and components perform under various environmental and usage conditions. For products with a digital interface, synthetic datasets can help in understanding how users interact with the product, enabling designers to optimize the user experience.

4.2. End-to-end datasets

Design research substantially relies on various natural language schemes, such as ontologies, controlled language descriptions, and documentation templates, to facilitate early phase design process and enhance design strategies (Siddharth et al., 2022). This is depicted in two smaller and lower double diamond processes shown in Figure 2 which shows that the applications that exist are only covering early design phases. The bigger double diamond process shows two expansions and narrowing down that happen during the design process. To understand better, the 4 steps of the traditional product development process are also visualized above in the same figure and several tasks and tools that traditionally are used in each step are written on each phase. Six categories of geometry-based datasets (reviewed in the previous section) are also depicted above this traditional PD process with G1 through G6. These datasets show most of the geometry-based datasets that exist in the literature are for detail design and testing and refinement phases of product development.

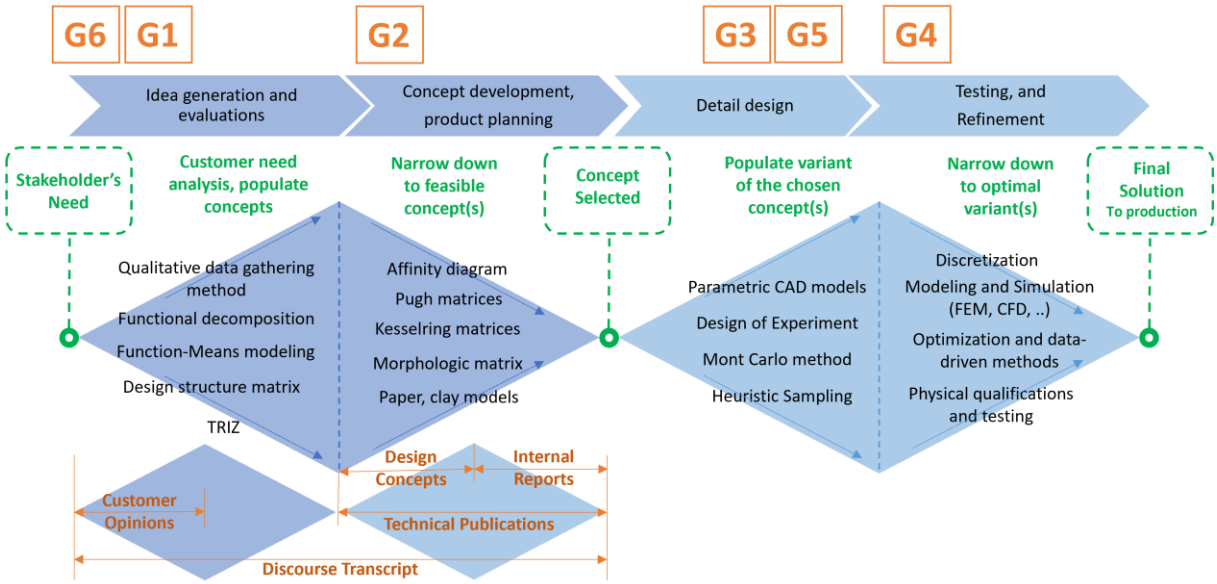

Figure 2. Identified design datasets categories (from Table 1) and NLP research in design (from (Siddharth et al., 2022)) in relation to two product development processes

The figure also demonstrates gaps for diverse and specialized datasets in NLP applications that is also raised by (Siddharth et al., 2022). Datasets should focus on text cleaning, term identification, and relation extraction in the early phases. The design community should focus on creating Domain-Specific Language Models, that demand datasets in text classification and sentiment analysis. These specialized models can be enhanced with detailed datasets for term identification and ontology construction in specific research areas or industries. Text Generation applications useful for example for concept studies, require comprehensive and varied content. This can be achieved by collaborative tagging benefits from diverse, user-generated content. This study also emphasizes standard datasets, addressing both creative and functional aspects.

The study reveals a notable gap between geometry-based datasets and their integration in the early design datasets. To bridge this gap, there is a need for datasets that include functionally decomposed products and functional annotations on geometries, fostering a stronger connection between a product's form and its intended function. Additionally, the incorporation of function-oriented design solutions, tailored for various products or specific industries, would be beneficial. Furthermore, the design community could greatly benefit from comprehensive end-to-end datasets. These datasets should

https://doi.org/10.1017/pds.2024.194 Published online by Cambridge University Press

encapsulate the entire design process, starting from customer needs—possibly derived from Customer Opinion analysis in NLP applications—through to the functional requirements of these needs, culminating in geometry-based design solutions directly linked to these functions. Such holistic datasets would enable a more integrated and functionally aligned approach to product design.

5. Can AI and GPT fill in the gap?

Research on GPT and design is very young and fast evolving, but it has already taken a fast pace. To facilitate bio-inspired design, recently a study used generative GPT models to automatically retrieve and map biological analogies for use in design problems (Zhu et al., 2023). The approach involved fine-tuning three types of design concept generators based on GPT-3 and using machine evaluators to assess the relevancy of these mappings. GPTs have also been used for early-stage design concept generation (Zhu and Luo, 2023) by transforming textual data into new design concepts. The approach is demonstrated through three tasks: domain knowledge synthesis, problem-driven synthesis, and analogy-driven synthesis. The study shows this method's effectiveness in generating novel and effective design concepts, highlighting its application in AI-driven design processes. Other have (Girotra et al., 2023) generated 200 design product ideas for under 50 dollars using GPT-4, and the same task is asked to be carried out by students in product design and innovation courses. The comparison showed not only ChatGPT is much faster and cheaper in generating ideas, but also the quality of the concepts is reported to be higher than human-generated concepts. In another attempt (Nelson et al., 2023) ChatGPT has been successfully used to assist CAD design by generating code that draws some simple functional microfluidic devices.

There are reports of unsuccessful attempts to use ChatGPT for mechanical engineering calculations (Tiro, 2023). The paper via several machine design examples, reports hallucinating formulas, selecting the wrong procedures to solve problems, and wrong mathematical calculations. A framework is proposed to assess ChatGPT's impact on generating innovative concepts in product design (Filippi, 2023). The study found that ChatGPT increased the quantity of concepts generated, but concepts were less useful compared to classic design methods. They report that GPT performed well in generating novel concepts, however, its impact on the variety of concepts was minimal.

It can be inferred from the published work on GPTs role, that AI and tools like GPT are capable of creating innovative concepts and ideation rather than quantitative work. Utilizing such capabilities can enable designers to create synthetic and end-to-end datasets proposed in this paper. AI algorithms can be trained to produce realistic and diverse datasets that accurately mimic real-world conditions. GPT, specifically, can be instrumental in generating realistic user interactions or textual data, such as customer reviews or product descriptions. This can aid in understanding user sentiment and preferences, which are crucial in product design. Additionally, AI can analyze these complex datasets to uncover insights that might not be apparent through traditional analysis methods and help further enhance product development processes.

The implementation of AI in generating data brings forth data that is not just varied in its parameters but is intricately designed to replicate the unpredictability of real-world user interactions. By integrating AI, we can synthesize, augment, and anonymize data. AI's ability to produce synthetic data allows for the creation of entirely new datasets that obey specific rules and patterns, reflecting realistic user details such as names and contact information without breaching confidentiality. Additionally, AI can enhance existing datasets through augmentation, adding complexity and variability to reflect real-world conditions, or through anonymization techniques that safeguard sensitive information while maintaining the integrity of the data.

Initially, it is crucial to meticulously define the requirements for the test data, considering aspects such as data types, structure, volume, and limitations. These criteria guide the selection of the most suitable AI methods and tools. Following this, validating the generated data against the set requirements confirms its accuracy and relevance. Moreover, managing the test data effectively ensures its continual relevance and accessibility. By adhering to these best practices, AI can be strategically used to generate datasets that are not only realistic and extensive but also respectful of privacy laws, thereby elevating the effectiveness of evaluations in product design and development.

https://doi.org/10.1017/pds.2024.194 Published online by Cambridge University Press

6. Conclusions

This paper underscores the role of datasets in advancing data-driven design research. Despite the diversity of existing datasets, gaps remain, particularly in the early design stages. Overcoming challenges in dataset creation, such as privacy, quality, and economic constraints, is crucial. The integration of AI and generative models like GPT shows promise in bridging these gaps, particularly in concept generation and ideation. However, limitations in quantitative analysis highlight the need for ongoing research and development. Embracing synthetic and end-to-end datasets is essential for a more holistic and integrated approach to product development, paving the way for a data-rich, innovative future in design research.

Acknowledgment

This work has been supported by the CHEOPS project, funded by the European Union under Grant Agreement Number 101004331.

References

Ajayi, K., Wei, X., Gryder, M., Shields, W., Wu, J., Jones, S.M., Kucer, M., Oyen, D., 2023. DeepPatent2: A Large-Scale Benchmarking Corpus for Technical Drawing Understanding. Sci Data 10, 772. https://doi.org/10.1038/s41597-023-02653-7

Arjomandi Rad, M., Cenanovic, M., Salomonsson, K., 2023. Image regression-based digital qualification for simulation-driven design processes, case study on curtain airbag. Journal of Engineering Design 34, 1–22. https://doi.org/10.1080/09544828.2022.2164440

Babu, S.S., Mourad, A.-H.I., Harib, K.H., Vijayavenkataraman, S., 2023. Recent developments in the application of machine-learning towards accelerated predictive multiscale design and additive manufacturing. Virtual and Physical Prototyping 18. https://doi.org/10.1080/17452759.2022.2141653

Bagazinski, N.J., Ahmed, F., 2023. Ship-D: Ship Hull Dataset for Design Optimization using Machine Learning.

Birkhofer, H., Lindemann, U., Weber, C., 2012. A View on Design: The German Perspective. Journal of Mechanical Design 134, 110301. https://doi.org/10.1115/1.4007847

Chan, Y.-C., Ahmed, F., Wang, L., Chen, W., 2021. METASET: Exploring Shape and Property Spaces for Data-Driven Metamaterials Design. Journal of Mechanical Design 143, 031707. https://doi.org/10.1115/1.4048629

Chang, A.X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., Xiao, J., Yi, L., Yu, F., 2015. ShapeNet: An Information-Rich 3D Model Repository. https://doi.org/10.48550/arXiv.1512.03012

Chen, W., Chiu, K., Fuge, M., 2019. Aerodynamic Design Optimization and Shape Exploration using Generative Adversarial Networks, in: AIAA Scitech 2019 Forum. Presented at the AIAA Scitech 2019 Forum, American Institute of Aeronautics and Astronautics, San Diego, California. https://doi.org/10.2514/6.2019-2351

Chiarello, F., Belingheri, P., Fantoni, G., 2021. Data science for engineering design: State of the art and future directions. Computers in Industry 129, 103447. https://doi.org/10.1016/j.compind.2021.103447

Curtis, B., Oliveira, V., 2023. Faker. URL https://github.com/faker-ruby/faker (accessed 11.15.23).

Du, X., Bilgen, O., Xu, H., 2021. Generating Pseudo-Data to Enhance the Performance of Classification-Based Engineering Design: A Preliminary Investigation. Presented at the ASME 2020 International Mechanical Engineering Congress and Exposition, American Society of Mechanical Engineers Digital Collection. https://doi.org/10.1115/IMECE2020-24634

Ejlli, D., 2022. Mockaroo: Generate Data To Practice With SQL And Python. Physics and Machine Learning. URL https://medium.com/physics-and-machine-learning/mockaroo-generate-data-to-practice-with-sql-and-python-7e581bc6d583 (accessed 11.12.23).

Filippi, S., 2023. Measuring the impact of ChatGPT on fostering concept generation in innovative product design. Electronics 12, 3535.

Girotra, K., Meincke, L., Terwiesch, C., Ulrich, K.T., 2023. Ideas are Dimes a Dozen: Large Language Models for Idea Generation in Innovation. https://doi.org/10.2139/ssrn.4526071

Goodman, E., 2014. Design and ethics in the era of big data. interactions 21, 22–24. https://doi.org/10.1145/2598902

Gorkovenko, K., Burnett, D.J., Thorp, J.K., Richards, D., Murray-Rust, D., 2020. Exploring The Future of Data-Driven Product Design, in: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. Presented at the CHI '20: CHI Conference on Human Factors in Computing Systems, ACM, Honolulu HI USA, pp. 1–14. https://doi.org/10.1145/3313831.3376560

Gryaditskaya, Y., Sypesteyn, M., Hoftijzer, J.W., Pont, S.C., Durand, F., Bousseau, A., 2019. OpenSketch: a richly-annotated dataset of product design sketches. ACM Trans. Graph. 38, 232–1.
Guillard, B., Remelli, E., Yvernay, P., Fua, P., 2021. Sketch2Mesh: Reconstructing and Editing 3D Shapes from Sketches, in: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Presented at the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), IEEE, Montreal, QC, Canada, pp. 13003–13012. https://doi.org/10.1109/ICCV48922.2021.01278
Han, W., Xiang, S., Liu, C., Wang, R., Feng, C., 2020. SPARE3D: A Dataset for SPAtial REasoning on Three-View Line Drawings, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Presented at the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Seattle, WA, USA, pp. 14678–14687. https://doi.org/10.1109/CVPR42600.2020.01470
Heyrani Nobari, A., Srivastava, A., Gutfreund, D., Ahmed, F., 2022. LINKS: A Dataset of a Hundred Million Planar Linkage Mechanisms for Data-Driven Kinematic Design. Presented at the ASME 2022 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, American Society of Mechanical Engineers Digital Collection. https://doi.org/10.1115/DETC2022-89798
Hoang, V.-N., Nguyen, N.-L., Tran, D.Q., Vu, Q.-V., Nguyen-Xuan, H., 2022. Data-driven geometry-based topology optimization. Struct Multidisc Optim 65, 69. https://doi.org/10.1007/s00158-022-03170-8
Jayanti, S., Kalyanaraman, Y., Iyer, N., Ramani, K., 2006. Developing an engineering shape benchmark for CAD models. Computer-Aided Design 38, 939–953. https://doi.org/10.1016/j.cad.2006.06.007
Kasturi, S., 2020. Some Aspects of Test Data Management Strategy, in: 2020 IEEE International Conference on Computing, Power and Communication Technologies (GUCON). IEEE, pp. 6–12.
Kim, S., Chi, H., Hu, X., Huang, Q., Ramani, K., 2020. A Large-Scale Annotated Mechanical Components Benchmark for Classification and Retrieval Tasks with Deep Neural Networks, in: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (Eds.), Computer Vision – ECCV 2020, Lecture Notes in Computer Science. Springer International Publishing, Cham, pp. 175–191. https://doi.org/10.1007/978-3-030-58523-5_11
Koch, P.N., Simpson, T.W., Allen, J.K., Mistree, F., 1999. Statistical Approximations for Multidisciplinary Design Optimization: The Problem of Size. Journal of Aircraft 36, 275–286. https://doi.org/10.2514/2.2435
Koch, S., Matveev, A., Jiang, Z., Williams, F., Artemov, A., Burnaev, E., Alexa, M., Zorin, D., Panozzo, D., 2019. ABC: A Big CAD Model Dataset for Geometric Deep Learning, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Presented at the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 9593–9603. https://doi.org/10.1109/CVPR.2019.00983
Lejeune, E., 2020. Mechanical MNIST: A benchmark dataset for mechanical metamodels. Extreme Mechanics Letters 36, 100659. https://doi.org/10.1016/j.eml.2020.100659
Málaga-Chuquitaype, C., 2022. Machine Learning in Structural Design: An Opinionated Review. Frontiers in Built Environment 8.
Manda, B., Dhayarkar, S., Mitheran, S., Viekash, V.K., Muthuganapathy, R., 2021. 'CADSketchNet' - An Annotated Sketch dataset for 3D CAD Model Retrieval with Deep Neural Networks. Computers & Graphics 99, 100–113. https://doi.org/10.1016/j.cag.2021.07.001
Mo, K., Zhu, S., Chang, A.X., Yi, L., Tripathi, S., Guibas, L.J., Su, H., 2019. PartNet: A Large-Scale Benchmark for Fine-Grained and Hierarchical Part-Level 3D Object Understanding, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Presented at the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Long Beach, CA, USA, pp. 909–918. https://doi.org/10.1109/CVPR.2019.00100
Moghaddam, M., Marion, T., Holtta-Otto, K., Fu, K., Olechowski, A., McComb, C. (Eds.), 2023. Special Issue: Emerging Technologies and Methods for Early-Stage Product Design and Development. Journal of Mechanical Design 145. https://doi.org/10.1115/1.4056744
Nelson, M.D., Goenner, B.L., Gale, B.K., 2023. Utilizing ChatGPT to assist CAD design for microfluidic devices. Lab on a Chip 23, 3778–3784. https://doi.org/10.1039/d3lc00518f
Panarotto, M., Isaksson, O., Habbassi, I., Cornu, N., 2022. Value-Based Development Connecting Engineering and Business: A Case on Electric Space Propulsion. IEEE Transactions on Engineering Management 69, 1650–1663. https://doi.org/10.1109/TEM.2020.3029677
Panchal, J.H., Fuge, M., Liu, Y., Missoum, S., Tucker, C. (Eds.), 2019. Special Issue: Machine Learning for Engineering Design. Journal of Mechanical Design 141. https://doi.org/10.1115/1.4044690
Philip Chen, C.L., Zhang, C.-Y., 2014. Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. Information Sciences 275, 314–347. https://doi.org/10.1016/j.ins.2014.01.015
Picard, C., Schiffmann, J., Ahmed, F., 2023. DATED: Guidelines for Creating Synthetic Datasets for Engineering Design Applications.
Rad, M.A., Salomonsson, K., Cenanovic, M., Balague, H., Raudberget, D., Stolt, R., 2022. Correlation-based feature extraction from computer-aided design, case study on curtain airbags design. Computers in Industry 138, 103634. https://doi.org/10.1016/j.compind.2022.103634

Ramnath, S., Haghighi, P., Kim, J.H., Detwiler, D., Berry, M., Shah, J.J., Aulig, N., Wollstadt, P., Menzel, S., 2019. Automatically Generating 60,000 CAD Variants for Big Data Applications. Presented at the ASME 2019 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, American Society of Mechanical Engineers Digital Collection. https://doi.org/10.1115/DETC2019-97378

Regenwetter, L., Curry, B., Ahmed, F., 2021. BIKED: A Dataset for Computational Bicycle Design with Machine Learning Benchmarks. https://doi.org/10.48550/arXiv.2103.05844

Regenwetter, L., Nobari, A.H., Ahmed, F., 2022. Deep Generative Models in Engineering Design: A Review. https://doi.org/10.48550/arXiv.2110.10863

Renear, A.H., Sacchi, S., Wickett, K.M., 2010. Definitions of dataset in the scientific and technical literature. Proceedings of the American Society for Information Science and Technology 47, 1–4. https://doi.org/10.1002/meet.14504701240

Shugrina, M., Li, C.-Y., Fidler, S., 2022. Neural Brushstroke Engine: Learning a Latent Style Space of Interactive Drawing Tools. ACM Trans. Graph. 41, 1–18. https://doi.org/10.1145/3550454.3555472

Siddharth, L., Blessing, L., Luo, J., 2022. Natural language processing in-and-for design research. Design Science 8, e21. https://doi.org/10.1017/dsj.2022.16

Song, B., Zhou, R., Ahmed, F., 2023. Multi-modal Machine Learning in Engineering Design: A Review and Future Directions. https://doi.org/10.48550/arXiv.2302.10909

Tiro, D., 2023. The Possibility of Applying ChatGPT (AI) for Calculations in Mechanical Engineering, in: Karabegovic, I., Kovačević, A., Mandzuka, S. (Eds.), New Technologies, Development and Application VI, Lecture Notes in Networks and Systems. Springer Nature Switzerland, Cham, pp. 313–320. https://doi.org/10.1007/978-3-031-31066-9_34

Ulrich, K.T., Eppinger, S.D., Yang, M.C., 2008. Product design and development. McGraw-Hill higher education Boston.

Wang, G.G., Shan, S., 2006. Review of Metamodeling Techniques in Support of Engineering Design Optimization. Journal of Mechanical Design 129, 370–380. https://doi.org/10.1115/1.2429697

Whalen, E., Beyene, A., Mueller, C., 2021. SimJEB: Simulated Jet Engine Bracket Dataset. https://doi.org/10.1111/cgf.14353

Willis, K.D.D., Pu, Y., Luo, J., Chu, H., Du, T., Lambourne, J.G., Solar-Lezama, A., Matusik, W., 2021. Fusion 360 gallery: a dataset and environment for programmatic CAD construction from human design sequences. ACM Trans. Graph. 40, 1–24. https://doi.org/10.1145/3450626.3459818

Wu, Z., Song, S., Khosla, A., Yu, F., Zhang, L., Tang, X., Xiao, J., 2015. 3d shapenets: A deep representation for volumetric shapes, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1912–1920.

Xu, P., Huang, Y., Yuan, T., Pang, K., Song, Y.-Z., Xiang, T., Hospedales, T.M., Ma, Z., Guo, J., 2018. SketchMate: Deep Hashing for Million-Scale Human Sketch Retrieval, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Presented at the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Salt Lake City, UT, USA, pp. 8090–8098. https://doi.org/10.1109/CVPR.2018.00844

Yang, M., Jiang, P., Zang, T., Liu, Y., 2023. Data-driven intelligent computational design for products: method, techniques, and applications. Journal of Computational Design and Engineering 10, 1561–1578. https://doi.org/10.1093/jcde/qwad070

Yüksel, N., Börklü, H.R., Sezer, H.K., Canyurt, O.E., 2023. Review of artificial intelligence applications in engineering design perspective. Engineering Applications of Artificial Intelligence 118, 105697. https://doi.org/10.1016/j.engappai.2022.105697

Zheng, L., Kumar, S., Kochmann, D.M., 2023. Unifying the design space of truss metamaterials by generative modeling.

Zhu, Q., Luo, J., 2023. Generative Transformers for Design Concept Generation. Journal of Computing and Information Science in Engineering 23. https://doi.org/10.1115/1.4056220

Zhu, Q., Zhang, X., Luo, J., 2023. Biologically Inspired Design Concept Generation Using Generative Pre-Trained Transformers. Journal of Mechanical Design 145. https://doi.org/10.1115/1.4056598