Automated Patent Landscaping

Sebastian Erhardt[1,*]

[1]*Max Planck Institute for Innovation and Competition, Marstallplatz 1, Munich, Bavaria, 80539, Germany*

Abstract

Patent Landscaping is a valuable instrument for many stakeholders, such as patent examiners, company decision-makers, researchers, and policymakers. They use this method to analyze the state-of-the-art, compare organizations' patenting activities, assess entire industries, or identify gaps in internal R&D activities. However, analyzing vast amounts of patent documents and aggregating and visualizing information is cumbersome and complex. The paper presents an innovative approach to automated patent landscaping by combining natural language processing models with approximate nearest neighbor search, dimensionality reduction, and clustering methods. This entire approach only uses the textual content of the underlying patents and does not use any additional meta-data, such as technology classes or citations.

Keywords

Patent Landscaping, Patent Portfolio Analysis, Machine Learning, Dimensionality Reduction, Clustering,

1. Introduction

The patent landscaping process involves analyzing and visualizing a collection of patents within a specific technological field, industry, or organization. The objective is to gain insights into the intellectual property landscape. Depending on the unit of analysis, the results can be utilized for various purposes.

The process is used by diverse stakeholders, including researchers, patent examiners, decision-makers, R&D managers, investors, and policymakers [1]. It is a crucial tool for various practical applications, helping organizations navigate the complex terrain of intellectual property. Patent landscaping informs strategies and decisions across multiple domains by providing valuable insights into the current state of technology, emerging trends, key players, and potential competitors. These insights are tailored to meet specific needs, such as guiding research directions, enhancing patent examination processes, making informed investment choices, and shaping organizational and policy frameworks. That helps stakeholders to manage and innovate adeptly within their respective fields.

One of the primary uses of patent landscaping is to assess technological progress and track innovation within specific fields. This involves identifying which organizations — be they companies, research institutions, or other entities — are actively working in an area, what technologies and industries they are targeting, and how technical problems are being solved. Additionally, understanding where patents are being filed and who the key players are helps stakeholders gauge the breadth and depth of activity in a particular domain.

Patent landscaping also plays a significant role in detecting potential infringements and assessing the validity of existing patents by examining their legal status. This provides stakeholders with a clear view of the competitive landscape and potential barriers to entry for new innovations. For businesses and investors, this information is critical for designing around existing technologies, identifying licensing and merger and acquisition targets, and reducing legal risks associated with intellectual property.

Furthermore, patents play a crucial role in helping organizations secure external financing. They can attract venture capital [2], serve as collateral for debt assignments [3], and even increase the valuation during initial public offerings (IPOs) [4]. However, investors require a comprehensive overview of these assets. They must also evaluate the competition and assess the value of these assets to identify potential risks [5].

Moreover, patent landscapes can significantly influence policy-making. Governments and international institutions use patent landscapes as a critical factor in developing science and technology policy [6]. For example, the OECD generates patent landscapes of different areas to map scientific and technological trends [1]. The World Intellectual Property Organization (WIPO) also publishes its patent landscapes. The organization is mandated to produce patent landscape reports[1] in areas of particular interest to developing and least developed countries, such as public health, food security, climate change, and the environment.

Organizations use this process internally for strategic research and technology transfer decisions. It helps stakeholders understand current trends and patterns in patenting activity and innovation, guiding informed decision-making processes. This includes optimizing internal R&D

5th Workshop on Patent Text Mining and Semantic Technologies (PatentSemTech) 2024

*Corresponding author.

sebastian.erhardt@ip.mpg.de (S. Erhardt)

[1] See https://www.wipo.int/patentscope/en/programs/patent_landscapes/ (retrieved 22.04.2024)

processes and establishing a comprehensive and well-informed IP strategy [1].

Furthermore, patent landscaping aids in exploring competitor activity and enhancing competition and market intelligence. It provides insights into where competitors invest resources and predicts the next products they will likely release. Of course, they are also constantly monitoring their competitors for possible infringements of their patents. They also look out for potential targets for M&A activities [7].

By presenting complex patent data visually, stakeholders can better comprehend and derive actionable insights based on empirical evidence. Visual representations such as charts, graphs, heat maps, network diagrams, and other visualizations are crucial in identifying patterns, relationships, and clusters within the patent dataset. These tools make it easier for decision-makers to grasp the intricacies of the intellectual property landscape at a glance.[2]

Interpreting these visual findings involves drawing conclusions about potential opportunities, gaps, and threats within the IP landscape. This analysis is vital for guiding decision-making, strategic planning, and fostering innovation efforts. Through effective visualization, patent landscaping not only aids in understanding but also significantly improves communication among stakeholders. This facilitates a more informed and collaborative approach to managing and capitalizing on intellectual property.

While patent landscaping is an invaluable tool, it is essential to acknowledge that it is also a complex and time-consuming process [5, 8, 9, 1, 10]. This complexity arises primarily because the detailed analysis and comparison of patent portfolios require a deep understanding of patent searching, analysis, and interpretation. Each of these tasks demands a high level of expertise, which many companies and organizations may not have in-house. Finally, the accuracy of these reports is also limited since these processes can be prone to human error since they rely on manual review and analysis of patent documents. This can lead to an inaccurate or incomplete analysis of the patent landscape [9, 8]. Moreover, much of the work in patent landscaping, including assessing vast arrays of data and identifying relevant patterns and insights, is done manually. However, since the number of analyzed patent documents can reach hundreds of thousands [11], it is impossible to even for the largest teams of experts.

Despite its complexities, the benefits of making better-informed decisions through patent landscaping are significant and multifaceted. Firstly, eliminating redundant research efforts helps reduce the costs associated with research and development. At the same time, it shortens the time needed for commercialization and ensures that resources are used more efficiently. Moreover, patent landscaping facilitates increased revenue opportunities through licensing. By identifying potential licensing opportunities, companies can monetize their patents beyond direct product sales, leveraging their intellectual property for broader financial gains. Additionally, patent landscaping provides researchers, policy-makers, companies, and investors with a better and clearer overview of the patent environment for specific technologies. This comprehensive understanding allows for strategic decision-making, offering insights into product differentiation and market positioning based on intellectual property analysis.

2. Related Work

There are simple off-the-shelf metadata visualization platforms like PatentsView that are used for patent landscaping. It was initiated in 2012 by the U.S. Patent and Trademark Office and is a comprehensive platform for visualizing, disseminating, and analyzing intellectual property data. It is tailored to support a diverse user base, including researchers, policymakers, and small business owners. It offers several essential tools for patent landscaping: visual analytics for exploring patent data, an API tool for data integration and advanced querying, and a data query builder for creating customized datasets.[3]

Trippe (2015), listed a vast array of analysis tool providers. Many of these tools use the technologies described in this section.

To better understand a variety of patent documents and associated technological information, previous research has primarily utilized the classification systems established by patent offices [12, 13, 14]. However, these classifications often lack the granularity necessary to fully represent the specific technologies involved [15, 16, 17, 18].

Determining patent similarity is a fundamental aspect of patent landscaping. This typically involves analyzing the text of patent documents using various text-mining algorithms. Initially, simple vectorized keyword extraction methods were employed to understand the relationships among claim elements [19].

Additionally, semantic text grammars have been used to delineate Subject Action Object (SAO) structures within patents [20, 21, 9, 22, 23, 24, 25].

Lexical databases, such as WordNet, have also played a role in mining patents for key technological concepts, enhancing the extraction of meaningful data [9, 23].

More recently, models that focus on semantic similarity leveraging word embeddings have been explored. Skripnikova et al. (2021) applied a pre-trained word2vec

[2] Source: https://www.wipo.int/patentscope/en/programs/patent_landscapes (retrieved 22.04.2024)

[3] Source: https://patentsview.org/what-is-patentsview (retrieved 22.04.2024)

model in combination with TF-IDF to create patent landscapes, utilizing dimensionality reduction and clustering techniques to analyze document relationships. Similarly, Abood and Feltenberger (2018) introduced an automated landscaping approach using patent metadata and word embeddings.

Erhardt et al. (2022) employed the SPECTER [27] model to generate document embeddings in combination with a comprehensive system for semantic searches. The system is tailored to patents and scientific publications.

Since SPECTER was only trained on scientific publications, Ghosh et al. (2024) further enhanced this field by presenting dedicated patent similarity models.

This paper describes an approach to fuse these advanced methodologies by Erhardt et al. and Ghosh et al., building upon the frameworks established by Abood and Feltenberger and Skripnikova et al. to refine automated patent landscaping techniques.

3. Methodology

This section introduces a new automated method for creating patent landscapes. This method significantly reduces many previously discussed constraints by utilizing human insights, semantic similarity, approximate nearest neighbor search, dimensionality reduction, and clustering. The most important aspect of this approach is that it only relies on text. No additional metadata is needed.

Logic Mill This approach relies heavily on the capabilities of *Logic Mill* [26], a software system wrapped around an extensive vector database. It is designed to find semantic similar documents across single or multiple domain-specific corpora. It employs state-of-the-art Natural Language Processing (NLP) techniques to create numerical representations of documents, utilizing a vast pre-trained language model for this purpose. The system specializes in analyzing scientific and patent documents and includes a database of over 200 million documents. Users can access Logic Mill through an Application Programming Interface (API) or a web interface. *Logic Mill* is regularly updated and can be adapted to include text corpora from various other fields. It is envisioned as a versatile tool for future research in the social sciences and beyond. The approach uses it to retrieve pre-computed embeddings and identify the nearest neighbors of reference documents.

3.1. Patent Data

The core of the process are patent documents, and especially patent text. Patent documents contain multiple text segments, such as a title, abstract, detailed description, and patent claims. *PATSTAT*[4], offered by the European Patent Office, has established itself as a leading source of patent intelligence and statistics. It allows for advanced statistical analysis of patent data, including bibliographical and legal events. This approach was used as a source to obtain the title and abstract. Furthermore, additional metadata is used as well. Another benefit of *PATSTAT* is a simple aggregation of organization names. In addition to the default metadata, Patstat provides aggregated and harmonized organization names linked by ownership relations to the patent documents. That is especially helpful in pinpointing patent documents with company names that are written in different styles or missing company designations.

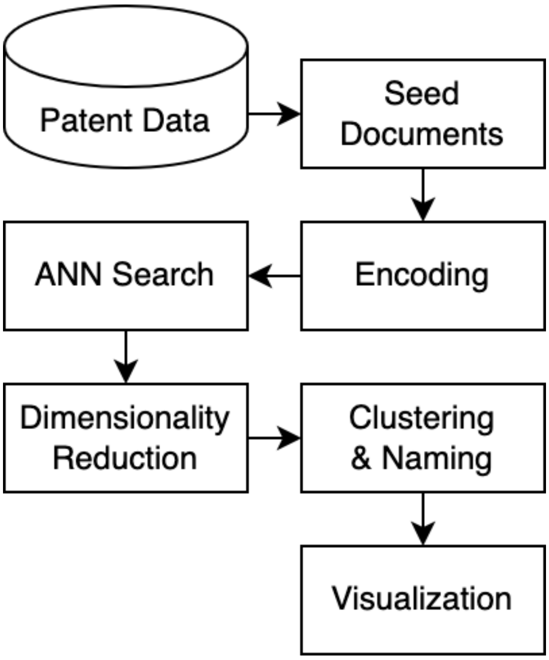

Figure 1: The steps of the approach

3.2. Seed Documents

Similar to the method introduced by Abood and Feltenberger (2018), users must select what are known as seed patents. Any mistakes made in choosing these seed patents will affect the entire resulting landscape. Therefore, it is crucial to ensure that the patents included in the seed set are relevant to the subject matter. The seed set should only contain patents that are relevant to the desired landscape and accurately represent the theme.

[4] See https://www.epo.org/searching-for-patents/business/patstat.html

Depending on the use case, the seed selection can comprise different documents. Based on the use case and the available metadata, end-users are able to gain insights in various contexts.

Hand Picked Technology Assessment For a technology assessment, one would start with the essential patents and then expand using related patents.

Firm-Level For a patent portfolio analysis of an organization's patent portfolio, one would start with the patent document of the focal organization. Along the way, competitors of the focal organization can be identified using the approximate nearest neighbor search.

CPC-Class Depending on the hierarchy, use the patent documents of a *Section*, *Class*, *Subclass*, *Group* or *Main Group* of the Cooperative Patent Classification (CPC) system and analyze the patents within your unit of analysis. Uncover differences and similarities between different hierarchical siblings.

Scientific Articles This text-based approach offers several benefits. First, it can bridge the gap between scientific publications and patents by using scientific papers as starting points. Additionally, this method can be applied to scholarly articles to create a comprehensive research landscape. Combining scientific and patent documents could also generate a useful tool for prior art exploration.

3.3. Encoding

All necessary documents need to be encoded before continuing. This can be done with any machine-learning document encoder model that can encode text into a dense numerical vector. In the case of patents, a reasonable candidate would be *SPECTER* Cohan et al. (2020. This machine-learning model encodes the title and the abstract of the patent into a 768-dimensional vector. Since the model was trained by leveraging the citation graph of scientific documents, it has learned the semantic similarities of related texts.

Other state-of-the-art models, like *SPECTER 2* [27], *PaECTER* or *Pat-SPECTER* of [28].

As stated before, we rely on the functionality of *Logic Mill* during this approach and make use of the precomputed embeddings for our data. The initial version of Logic Mill uses *SPECTER* [27], while the current version uses *Pat-SPECTER* [28].

3.4. ANN Search

A vector search database is needed to automatically find similar semantic documents. In this database, all necessary encoded documents are stored. According to a similarity metric (e.g., Cosine Similarity see Equation 1 or Euclidean Distance / L2 Distance see Equation 2), these documents are indexed, and the database is able to provide the approximate nearest neighbors in high-dimensional space according to these metrics. After all documents are in the database, users feed the encoded seed query documents to the database and retrieve the closest semantically similar documents. They can specify the number of nearest neighbors that should be included for each reference document. Otherwise, they also can specify a cut-off/threshold value. If the similarity score is below this value, the patent will not be added to the landscaping process.

In the case of *Logic Mill*, these representations are stored in a vector search database called *Elastic Search*[5]. It uses the approximate nearest neighbor algorithm HNSW [29] to retrieve semantically similar documents. This approach uses the Logic Mill API to retrieve the closest neighbors based on the IDs of the seed documents.

3.5. Dimensionality Reduction

Since the numerical representations of the encoded patent documents generated by *SPECTER* / *Pat-SPECTER* is a vector of 768 dimensions or 1024 by *PatSPECTER*, the dimensions have to be reduced since they are impossible to visualize and interpret for humans. This can be done through a process called dimensionality reduction. There are various methods, such as the linear variants *Principal Component Analysis* (PCA) [30] or *t-Distributed Stochastic Neighbor Embedding* (t-SNE) [31], as well as the non-linear variants such as *Uniform Manifold Approximation and Projection* (UMAP) [32]. The objective of these methods is to increase interpretability while at the same time minimizing the loss of information [33, 34].

For this approach, *UMAP* was chosen since it offers multiple advantages, as seen in Table 1. On the one hand, it can preserve the data's global and local structures. In addition, it can work with non-linear data, and finally, it is fast.

	PCA [30]	**tSNE** [31]	**UMAP** [32]
Non-linear data	No	No	Yes
Local structure	Yes	Yes	Yes
Global structure	No	Yes	Yes
Speed	Very Fast	Slow	Fast

Table 1
Overview of dimensionality reduction algorithms.

During the dimensionality reduction process of *UMAP*,

[5] See https://www.elastic.co/

various similarity measures can be selected to identify the most similar neighbors in the high-dimensional space. In this approach, cosine similarity and Euclidean distance can be used.

Cosine Similarity

$$\cos(\mathbf{a},\mathbf{b}) = \frac{\mathbf{a}\cdot\mathbf{b}}{\|\mathbf{a}\|\|\mathbf{b}\|} = \frac{\sum_{i=1}^{n}\mathbf{a}_i\mathbf{b}_i}{\sqrt{\sum_{i=1}^{n}(\mathbf{a}_i)^2}\sqrt{\sum_{i=1}^{n}(\mathbf{b}_i)^2}} \tag{1}$$

Euclidean Distance / L2 Distance

$$l2(\mathbf{a},\mathbf{b}) = \|\mathbf{b}-\mathbf{a}\|_2 = \sqrt{\sum_{i=1}^{n}(\mathbf{b}_i-\mathbf{a}_i)^2} \tag{2}$$

There are additional algorithm parameters that can be fine-tuned for every case. Initially, the default settings are selected.

The outcome of this reduction is the transformation from a $768/1024$ dimensional representation into a 2 or 3 dimensional one.

3.6. Clustering

Using the resulting 2/3-dimensional data structure from the previous dimensionality reduction step, a clustering algorithm is applied. The *HDBSCAN* [35] algorithm was selected for this approach. As seen in Table 2, its parameter *minimum amount of cluster members* is very intuitive, and the modest speed is negligible. In comparison, *K-Means* [36] would not be feasible here since it is not clear in advance how many clusters are going to be needed. Since it requires some domain knowledge to form the neighborhood parameter of *DBSCAN* [37], it was not a feasible option for a generic and automated end-to-end approach. Furthermore, Grootendorst (2022) showed promising results by combining *HDBSCAN* [35] with *UMAP* [32]. It also generates better results if the clustering is done after the dimensionality reduction [38]. The outcome of this step is a cluster ID label attached to each landscape document.

	K-Means [36]	**DBSCAN** [37]	**HDBSCAN** [35]
Type	Centroid Based	Density Based	Density / Hierarchical
Speed	Very Fast	Modest	Modest
Param.	Number of Clusters	Neighborhood	Number of Cluster Members

Table 2
Overview of clustering algorithms.

Additionally, the clustering algorithm includes parameters that can be adjusted after the user reviews an initial visualization. These settings vary depending on the specific use case (e.g., the number of data points and the density of certain areas). Initially, the default settings have been applied.

3.7. Cluster Naming

Based on the different clusters and the respective text of the patent documents, the approach uses the *c-TF-IDF* [38] algorithm to generate cluster names. This is also in line with Grootendorst (2022). Here, the traditional document-level textitTF-IDF is transformed to function effectively on a cluster-specific basis, which is essential for distinguishing the unique characteristics of each cluster. This approach, known as *c-TF-IDF*, modifies the standard method to reflect the differences between clusters better.

Initially, each cluster is treated as a singular document, aggregating the frequencies of each word within that cluster. This step forms the basis for a class-specific term frequency. An L1 normalization is applied to normalize these frequencies, ensuring that variations in cluster sizes do not skew the data.

Following this, the inverse document frequency is computed by taking the logarithm of the quotient of the average word count per cluster and the occurrence of each word across all clusters, incremented by one to maintain non-negative values. This calculation yields the modified IDF values.

c-TF-IDF

$$\mathbf{W}_{x,c} = \|\mathbf{tf}_{x,c}\| \times \log\left(1+\frac{\mathbf{A}}{\mathbf{f}_x}\right) \tag{3}$$

where $\mathbf{tf}_{x,c}$ is the frequency of word x in class c, $\mathbf{f}_x$ is the frequency of word x across all classes, and $\mathbf{A}$ is the average number of words per class.

Finally, these two metrics—the class-based term frequency and the adjusted inverse document frequency—are multiplied to derive a significance score for each word within a cluster. This method deviates from the conventional TF-IDF to provide a more tailored and effective topic representation [38].

In this approach, the *c-TF-IDF* aggregates the most common words per cluster and generates a fictional cluster name based on the, e.g., top 5 most common words.

Such cluster names would look like this:

- als neurodegenerative neurons disease mns
- manufacturing material printing materials 3d
- tumor nitrite taxol cancer lmp

3.8. Visualization

This approach visualizes traditional charts and graphs using the Plotly[6] library. This library can be used by Python, R, and JavaScript. It allows for interactive, web-based visualizations, including line plots, scatter plots, bar charts, and more. Plotly uses a declarative syntax, which makes it easy to create and customize 2D/3D plots. It also supports a wide range of chart types and customization options. Additionally, Plotly allows for the creation of interactive dashboards, which organize and present multiple plots in one place. Furthermore, the plots can be embedded into web pages and Jupyter notebooks. It is an open-source library that can be used in commercial and non-commercial applications [39].

A 3D/2D scatterplot is used to visualize the landscapes. Here, each dot is a patent document. The x and y (or z) coordinates result from the dimensionality reduction process.

For the visualization of clusters, the so-called convex hull was computed. The convex hull is commonly used in computational geometry and computer graphics to create a shape that encloses a group of points. In other words, the smallest "convex" shape can be drawn around a set of points such that all the points are on or inside the shape [40].

This approach uses the Virtanen et al. (2020) implementation of the Quickhull algorithm [40]. The input to the algorithm is the set of points of each cluster. The output is a polygon of the outermost points.

4. Results

In this section, two use cases are presented to demonstrate the approach.

4.1. Patent Landscaping - mRNA

During the COVID-19 pandemic, messenger RNA (mRNA) vaccines have attracted considerable attention. BioNTech and Moderna vaccines were the first mRNA vaccines to be approved by a drug regulatory agency worldwide. Intensive research had been conducted on mRNA technology for many years prior to the development of the COVID-19 vaccines[7].

The underlying technology used to develop such vaccines can be protected by patents. In an attempt to demonstrate the complexity involved in IP protections and licensing deals surrounding COVID-19 vaccine technology, Gaviria and Kilic (2021) developed a preliminary patent network analysis. Li et al. (2022) characterized the patent landscape of mRNA vaccines and analyzed 1852 patent families.

Objective This use case tries to generate a patent landscape analysis of mRNA patenting organizations. These findings are then compare to the studies of Gaviria and Kilic (2021) and Li et al. (2022).

Data BioNTech, CureVac, and Moderna have all developed mRNA-based vaccine candidates for COVID-19. Gaviria and Kilic (2021) also display Acuitas and Arbutus. The identification of the companies was made using the `han_name` column in the `tls206_person` table of Patstat. For the search, the term `BIONTECH, CUREVAC, MODERNATX, ACUITAS, ARBUTUS` were used. Only patents from the EPO were considered. No attention was given to whether the patents were still active. All patent documents (A1, A2, B1, etc.) were used. These initial documents represent the so-called seed documents in this approach.

Landscape Seven hundred fifty patent documents were retrieved from Patstat by using the organization search terms. The initial step is to obtain the numerical representation of these documents generated by SPECTER [27]. These were retrieved using the API of the Logic Mill system [26]. The embeddings were then used for dimensionality reduction and clustering. Finally, the convex hull of the cluster was calculated. Cluster Names were generated using the C-TF-IDF algorithm. An interactive visualization can be seen in Figure 2. These visualizations make use of 3 dimensions, and the clusters are indicated as transparent hulls around the dots. Each dot represents a patent document. The color indicated the respective companies. Users can now interactively explore the patent landscape based on the patents of these organizations.

Expand In the second step, the competitors of the selected companies were retrieved to generate the patent landscape of the mRNA industry. The already encoded patent documents of the initial organizations (BIONTECH, CUREVAC, MODERNATX, ACUITAS, ARBUTUs) were used as seed documents. The embeddings of the documents were used in combination with the Logic Mill API to retrieve the closest neighbors. For each reference document, the closest ten documents were retrieved. The result set was then used to obtain the organizations. Finally, the results are aggregated and counted based on the organization's name. The top 50 results can be found in Appendix 3. The `han_names` were not resolved to an organization level. There are multiple legal entities that belong to the same organization. This can be seen in the histogram in Figure 3.

[6] See https://plotly.com/

[7] See https://www.pei.de/DE/newsroom/hp-meldungen/2022/220221-covid-19-pandemie-impfstoffe-im-fokus.html?nn=169730 (retrieved 22.04.2024)

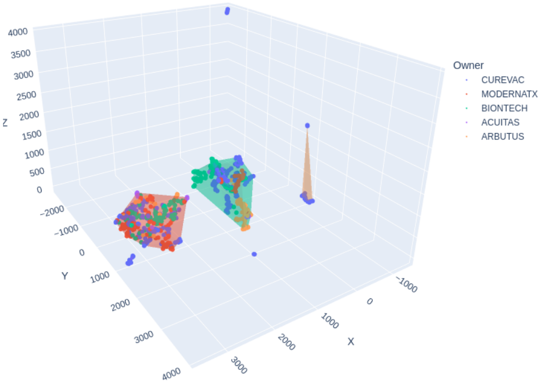

Figure 2: An interactive Plotly 3D scatter plot visualization of companies in the mRNA domain. Each dot represents a patent document. The color indicates the organization. Colored hulls surround patent clusters. Cluster Names are visible by hovering over the clusters.

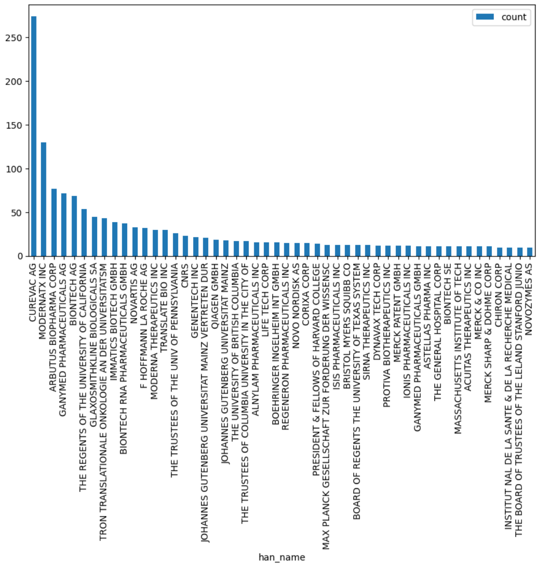

Figure 3: Histogram of automatically identified organizations by this approach. The number of documents was aggregated by the organization name provided by Patstat.

Comparing the automatically retrieved organizations with the work of Gaviria and Kilic 2021, overlaps can be identified. UPenn (University of Pennsylvania) and UBC (University of British Columbia) could be identified automatically.

Comparing the results with Li et al. 2022, additional overlaps could be found. Here, Merck & Co., GlaxoSmithKline, Boehringer, Roche, University of California, and TRON Translational Oncology Mainz were also represented in the patent landscape.

On the other hand, important mRNA organizations were missing. Sanofi, Enanta Pharmaceuticals, Evelo Biosciences, the United States Department of Health and Human Services, and the Chinese Academy of Agricultural Sciences were not present. This could be due to the fact that only patent documents of the EPO were analyzed. The analysis did not include the USPTO, WIPO, and the Chinese Patent Office.

Interpretation This use case shows the capabilities of the approach to automatically generate patent landscapes. Only initial reference organizations were selected. All other competitors have been identified automatically. This was done by only relying on the encoded patent text. With the help of an approximate nearest neighbor search in a vector search database.

4.2. Patent Landscaping - Quantum Computing

The latest technological advancements have shed light on the capabilities of quantum technology across various fields. Among them are simulation, computation, and communication. Quantum computers use superconducting qubit-based programmable processors. They can compute tasks in minutes, whereas state-of-the-art classical supercomputers would take approximately tens of thousands of years [44].

In their report, Aboy et al. (2022) produce a patent landscape of quantum technologies over the last 20 years. They evaluate patenting and strategies, key owners, dominant portfolios, and geographic distribution of patent activity, among other factors.

Objective This use case uses a single reference company in the quantum technologies domain to generate a patent landscape.

Data Based on the available information on the competitor landscape in the quantum domain by Aboy et al. (2022), `Rigetti` was selected as a reference organization.

The company was identified using the `han_name` column in the `tls206_person` table of Patstat. Only patents from the EPO were considered. No attention was given to whether the patents were still active. All patent documents (A1, A2, B1, etc.) were used. The `han_names` were not aggregated. The seed documents, in this case, are also all patent documents of the company `Rigetti`.

Analysis The initial response of Patstat was 23 patent documents. The numerical representation was obtained using the Logic Mill System [26]. The approximate nearest neighbors of the documents were retrieved using another API endpoint. The documents were then resolved back to the owner. Afterward, they are aggregated and counted.

The top 20 results can be seen in Figure 4. The top 50 results can be found in Appendix 4. The results of Aboy et al. (2022) can be found in the Appendix 4.

The overlap between the result generated by the approach and Aboy et al. (2022) is large. IBM, D-Wave, Northrop Grumman, Toshiba, MIT, Intel, Alphabet (Google), Honeywell, HP, Hitachi, Samsung, etc., were all in the top 50 results.

Furthermore, Lockheed Martin CORP (rank 80) QUBITEKK (rank 88) New South Innovations (rank 66) Raytheon Technologies (rank 52) were still within the top 100.

Only Bank of America, Seagate, and Michigan State University could not be found.

A reason could be that Patstat does not aggregate these documents correctly. Companies and their subsidiaries might not be aggregated. Furthermore, [45] used patent documents of the last 20 years of the USPTO and EPO.

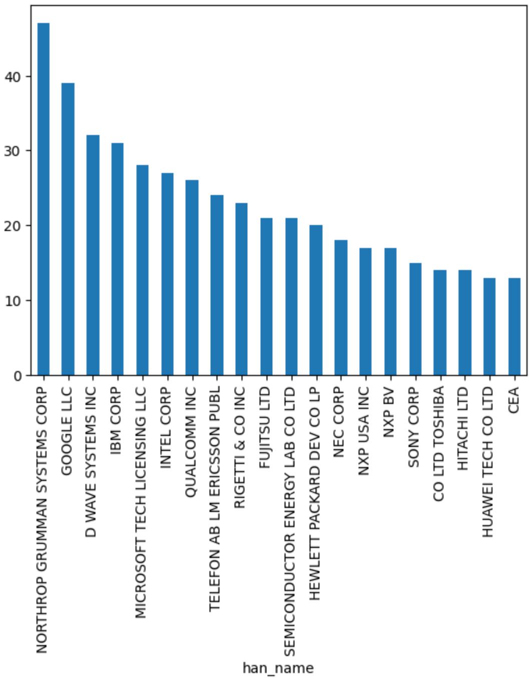

Figure 4: Amount of documents of competitors identified by the approximated nearest neighbors search based on patents of Rigetti.

Interpretation It was shown that the approach is capable of identifying competitors for a patent landscape. This was done only by using raw patent text, a document encoder, and a search database. The approximated nearest neighbor search returned competitors of Rigetti. These could be matched with the organizations presented by Aboy et al. 2022 in their patent landscape analysis.

5. Conclusions

The patent landscaping method outlined in this paper provides an automated and scalable solution, significantly reducing the need for human intervention. This approach has various benefits for researchers, companies, policymakers, and investors who rely on patent landscape analyses but are often deterred by their associated costs and efforts. The approach is notably fast, adaptable, capable of integrating new advancements and accommodating future research.

However, the approach has clear limitations. As it is based on seed documents, a careful selection is necessary. It is also challenging to determine in advance the right amount of neighbors or a threshold value for the approximate nearest neighbors of the seed documents. Additionally, the settings for dimensionality reduction and clustering might also depend on the use case and the number of documents.

Furthermore, this study is also limited in the evaluation of the approach. It is mostly episodic and limited in scope since the comparison relies only on EP data. A comparison with a gold-standard dataset is needed to compare the results of the automated approach with a human-generated one. It also lacks a comprehensive user study of the interface and visualization components.

Future research could explore how the latest generative AI models might be able to extend this approach further. With this technology, cluster naming can be improved greatly. It might also be possible to start the approach without any seed documents and start with a prompt. *Please create a patent landscape of mRNA vaccines against COVID-19.*

References

[1] A. Trippe, Guidelines for Preparing Patent Landscape Reports, World Intellectual Property Organization (WIPO) (2015).

[2] C. Häussler, D. Harhoff, E. Mueller, To Be Financed or Not… - The Role of Patents for Venture Capital-Financing, 2012. URL: https://papers.ssrn.com/abstract=1393725. doi:10.2139/ssrn.1393725.

[3] B. Amable, J.-B. Chatelain, K. Ralf, Patents as Collateral, 2010. URL: https://papers.ssrn.com/abstract=2227111. doi:10.2139/ssrn.2227111.

[4] M. B. Heeley, S. F. Matusik, N. Jain, Innovation, Appropriability, And The Underpricing Of Initial Public Offerings, Academy of Management Journal 50 (2007) 209–225. URL: https://journals.aom.org/doi/10.5465/amj.2007.24162388. doi:10.5465/amj.2007.24162388, publisher: Academy of Management.

[5] G. Littmann-Hilmer, M. Kuckartz, SME tailor-designed patent portfolio analysis, World Patent Information 31 (2009) 273–277. URL: https://www.sciencedirect.com/science/article/pii/S0172219008001762. doi:10.1016/j.wpi.2008.12.003.

[6] T. Bubela, E. R. Gold, G. D. Graff, D. R. Cahoy, D. Nicol, D. Castle, Patent landscaping for life sciences innovation: toward consistent and transparent practices, Nature Biotechnology 31 (2013) 202–206. URL: https://www.nature.com/articles/nbt.2521. doi:10.1038/nbt.2521, number: 3 Publisher: Nature Publishing Group.

[7] H. Ernst, Patent applications and subsequent changes of performance: evidence from time-series cross-section analyses on the firm level, Research Policy 30 (2001) 143–157. URL: https://www.sciencedirect.com/science/article/pii/S0048733399000980. doi:10.1016/S0048-7333(99)00098-0.

[8] M.-J. Shih, D.-R. Liu, M.-L. Hsu, Discovering competitive intelligence by mining changes in patent trends, Expert Systems with Applications 37 (2010) 2882–2890. URL: https://www.sciencedirect.com/science/article/pii/S0957417409007829. doi:10.1016/j.eswa.2009.09.001.

[9] H. Park, J. Yoon, K. Kim, Identifying patent infringement using SAO based semantic technological similarities, Scientometrics 90 (2012) 515–529. URL: https://doi.org/10.1007/s11192-011-0522-7. doi:10.1007/s11192-011-0522-7.

[10] A. Abood, D. Feltenberger, Automated patent landscaping, Artificial Intelligence and Law 26 (2018) 103–125. URL: https://doi.org/10.1007/s10506-018-9222-4. doi:10.1007/s10506-018-9222-4.

[11] T. Skripnikova, H. Aras, A. Weißhaar, S. Blank, H.-P. Zorn, Semantic views — Interactive hierarchical exploration for patent landscaping, World Patent Information 65 (2021) 102043. URL: https://www.sciencedirect.com/science/article/pii/S0172219021000247. doi:10.1016/j.wpi.2021.102043.

[12] A. B. Jaffe, Characterizing the "technological position" of firms, with application to quantifying technological opportunity and research spillovers, Research Policy 18 (1989) 87–97. URL: https://www.sciencedirect.com/science/article/pii/0048733389900073. doi:10.1016/0048-7333(89)90007-3.

[13] L. Rosenkopf, P. Almeida, Overcoming Local Search through Alliances and Mobility, Management Science 49 (2003) 751–766. URL: https://www.jstor.org/stable/4134022, publisher: INFORMS.

[14] N. Bloom, M. Schankerman, J. Van Reenen, Identifying Technology Spillovers and Product Market Rivalry, Econometrica 81 (2013) 1347–1393. URL: https://www.jstor.org/stable/23524180, publisher: [Wiley, Econometric Society].

[15] P. Thompson, M. Fox-Kean, Patent Citations and the Geography of Knowledge Spillovers: A Reassessment, American Economic Review 95 (2005) 450–460. URL: https://www.aeaweb.org/articles?id=10.1257/0002828053828509. doi:10.1257/0002828053828509.

[16] S. Arts, B. Cassiman, J. C. Gomez, Text matching to measure patent similarity, Strategic Management Journal 39 (2018) 62–84. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/smj.2699. doi:10.1002/smj.2699.

[17] C. Righi, T. Simcoe, Patent examiner specialization, Research Policy 48 (2019) 137–148. URL: https://www.sciencedirect.com/science/article/pii/S0048733318301902. doi:10.1016/j.respol.2018.08.003.

[18] S. Arts, N. Melluso, R. Veugelers, Beyond Citations: Text-Based Metrics for Assessing Novelty and its Impact in Scientific Publications, 2023. URL: http://arxiv.org/abs/2309.16437. doi:10.48550/arXiv.2309.16437, arXiv:2309.16437 [econ, q-fin].

[19] S. Lee, B. Yoon, Y. Park, An approach to discovering new technology opportunities: Keyword-based patent map approach, Technovation 29 (2009) 481–497. URL: https://www.sciencedirect.com/science/article/pii/S0166497208001326. doi:10.1016/j.technovation.2008.10.006.

[20] I. Bergmann, D. Butzke, L. Walter, J. P. Fuerste, M. G. Moehrle, V. Erdmann, Evaluating the Risk of Patent Infringement by Means of Semantic Patent Analysis: The Case of DNA Chips, 2008. URL: https://papers.ssrn.com/abstract=1288313. doi:10.1111/j.1467-9310.2008.00533.x.

[21] J. M. Gerken, M. G. Moehrle, A new instrument for technology monitoring: novelty in patents measured by semantic patent analysis, Scientometrics 91 (2012) 645–670. URL: https://doi.org/10.1007/s11192-012-0635-7. doi:10.1007/s11192-012-0635-7.

[22] S. Choi, H. Park, D. Kang, J. Y. Lee, K. Kim, An SAO-based text mining approach to building a technology tree for technology planning, Expert Systems with Applications 39 (2012) 11443–11455. URL: https://www.sciencedirect.com/science/article/pii/S0957417412006112. doi:10.1016/j.eswa.2012.04.014.

[23] H. Park, J. J. Ree, K. Kim, Identification of promising patents for technology transfers using TRIZ evolution trends, Expert Systems with Applications

40 (2013) 736–743. URL: https://www.sciencedirect.com/science/article/pii/S0957417412009694. doi:10.1016/j.eswa.2012.08.008.

[24] J. Yoon, H. Park, K. Kim, Identifying technological competition trends for R&D planning using dynamic patent maps: SAO-based content analysis, Scientometrics 94 (2013) 313–331. URL: https://doi.org/10.1007/s11192-012-0830-6. doi:10.1007/s11192-012-0830-6.

[25] H. Park, K. Kim, S. Choi, J. Yoon, A patent intelligence system for strategic technology planning, Expert Systems with Applications 40 (2013) 2373–2390. URL: https://www.sciencedirect.com/science/article/pii/S0957417412012092. doi:10.1016/j.eswa.2012.10.073.

[26] S. Erhardt, M. Ghosh, E. Buunk, M. E. Rose, D. Harhoff, Logic Mill – A Knowledge Navigation System, 2022. URL: http://arxiv.org/abs/2301.00200. doi:10.48550/arXiv.2301.00200, arXiv:2301.00200 [cs].

[27] A. Cohan, S. Feldman, I. Beltagy, D. Downey, D. Weld, SPECTER: Document-level Representation Learning using Citation-informed Transformers, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 2270–2282. URL: https://aclanthology.org/2020.acl-main.207. doi:10.18653/v1/2020.acl-main.207.

[28] M. Ghosh, S. Erhardt, M. E. Rose, E. Buunk, D. Harhoff, PaECTER: Patent-level Representation Learning using Citation-informed Transformers, 2024. URL: http://arxiv.org/abs/2402.19411. doi:10.48550/arXiv.2402.19411, arXiv:2402.19411 [cs].

[29] Y. A. Malkov, D. A. Yashunin, Efficient and robust approximate nearest neighbor search using Hierarchical Navigable Small World graphs, 2018. URL: http://arxiv.org/abs/1603.09320. doi:10.48550/arXiv.1603.09320, arXiv:1603.09320 [cs].

[30] I. T. Jolliffe, Principal Component Analysis, Springer Series in Statistics, Springer-Verlag, New York, 2002. URL: http://link.springer.com/10.1007/b98835. doi:10.1007/b98835.

[31] L. van der Maaten, G. Hinton, Visualizing data using t-SNE, Journal of Machine Learning Research 9 (2008) 2579–2605. URL: http://jmlr.org/papers/v9/vandermaaten08a.html.

[32] L. McInnes, J. Healy, UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, ArXiv (2018).

[33] I. T. Jolliffe, J. Cadima, Principal component analysis: a review and recent developments, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 374 (2016) 20150202. URL: https://royalsocietypublishing.org/doi/10.1098/rsta.2015.0202. doi:10.1098/rsta.2015.0202, publisher: Royal Society.

[34] R. Xiang, W. Wang, L. Yang, S. Wang, C. Xu, X. Chen, A Comparison for Dimensionality Reduction Methods of Single-Cell RNA-seq Data, Frontiers in Genetics 12 (2021). URL: https://www.frontiersin.org/articles/10.3389/fgene.2021.646936.

[35] L. McInnes, J. Healy, S. Astels, hdbscan: Hierarchical density based clustering, Journal of Open Source Software 2 (2017) 205. URL: https://joss.theoj.org/papers/10.21105/joss.00205. doi:10.21105/joss.00205.

[36] J. MacQueen, Some methods for classification and analysis of multivariate observations, Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics 5.1 (1967) 281–298. URL: https://projecteuclid.org/ebooks/berkeley-symposium-on-mathematical-statistics-and-probability/Proceedings-of-the-Fifth-Berkeley-Symposium-on-Mathematical-Statistics-and/chapter/Some-methods-for-classification-and-analysis-of-multivariate-observations/bsmsp/1200512992, publisher: University of California Press.

[37] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, A density-based algorithm for discovering clusters in large spatial databases with noise, in: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, KDD'96, AAAI Press, Portland, Oregon, 1996, pp. 226–231.

[38] M. Grootendorst, BERTopic: Neural topic modeling with a class-based TF-IDF procedure, 2022. URL: http://arxiv.org/abs/2203.05794. doi:10.48550/arXiv.2203.05794, arXiv:2203.05794 [cs].

[39] I. Plotly Technologies, Collaborative data science, 2015. URL: https://plot.ly, place: Montreal, QC Publisher: Plotly Technologies Inc.

[40] C. B. Barber, D. P. Dobkin, H. Huhdanpaa, The quickhull algorithm for convex hulls, ACM Transactions on Mathematical Software 22 (1996) 469–483. URL: https://doi.org/10.1145/235815.235821. doi:10.1145/235815.235821.

[41] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, I. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, SciPy 1.0: fundamental algorithms for scientific computing in Python, Nature Methods 17 (2020) 261–272. URL: https://www.nature.com/articles/s41592-019-0686-2. doi:10.1038/s41592-019-0686-2, number: 3

Publisher: Nature Publishing Group.

[42] M. Gaviria, B. Kilic, A network analysis of COVID-19 mRNA vaccine patents, Nature Biotechnology 39 (2021) 546–548. URL: https://www.nature.com/articles/s41587-021-00912-9. doi:10.1038/s41587-021-00912-9, number: 5 Publisher: Nature Publishing Group.

[43] M. Li, J. Ren, X. Si, Z. Sun, P. Wang, X. Zhang, K. Liu, B. Wei, The global mRNA vaccine patent landscape, Human Vaccines & Immunotherapeutics 18 (2022) 2095837. URL: https://doi.org/10.1080/21645515.2022.2095837. doi:10.1080/21645515.2022.2095837.

[44] F. Arute, K. Arya, R. Babbush, D. Bacon, J. C. Bardin, R. Barends, R. Biswas, S. Boixo, F. G. S. L. Brandao, D. A. Buell, B. Burkett, Y. Chen, Z. Chen, B. Chiaro, R. Collins, W. Courtney, A. Dunsworth, E. Farhi, B. Foxen, A. Fowler, C. Gidney, M. Giustina, R. Graff, K. Guerin, S. Habegger, M. P. Harrigan, M. J. Hartmann, A. Ho, M. Hoffmann, T. Huang, T. S. Humble, S. V. Isakov, E. Jeffrey, Z. Jiang, D. Kafri, K. Kechedzhi, J. Kelly, P. V. Klimov, S. Knysh, A. Korotkov, F. Kostritsa, D. Landhuis, M. Lindmark, E. Lucero, D. Lyakh, S. Mandrà, J. R. McClean, M. McEwen, A. Megrant, X. Mi, K. Michielsen, M. Mohseni, J. Mutus, O. Naaman, M. Neeley, C. Neill, M. Y. Niu, E. Ostby, A. Petukhov, J. C. Platt, C. Quintana, E. G. Rieffel, P. Roushan, N. C. Rubin, D. Sank, K. J. Satzinger, V. Smelyanskiy, K. J. Sung, M. D. Trevithick, A. Vainsencher, B. Villalonga, T. White, Z. J. Yao, P. Yeh, A. Zalcman, H. Neven, J. M. Martinis, Quantum supremacy using a programmable superconducting processor, Nature 574 (2019) 505–510. URL: https://www.nature.com/articles/s41586-019-1666-5. doi:10.1038/s41586-019-1666-5, number: 7779 Publisher: Nature Publishing Group.

[45] M. Aboy, T. Minssen, M. Kop, Mapping the Patent Landscape of Quantum Technologies: Patenting Trends, Innovation and Policy Implications, IIC - International Review of Intellectual Property and Competition Law 53 (2022) 853–882. URL: https://doi.org/10.1007/s40319-022-01209-3. doi:10.1007/s40319-022-01209-3.

A. Data

A.1. mRNA

Name	**Amount of docs**
CUREVAC AG	274
MODERNATX INC	130
ARBUTUS BIOPHARMA CORP	77
GANYMED PHARMACEUTICALS AG	72
BIONTECH AG	69
UNIVERSITY OF CALIFORNIA	54
GLAXOSMITHKLINE BIOLOGICALS SA	45
TRON - Universität Mainz	43
IMMATICS BIOTECH GMBH	39
BIONTECH RNA PHARMACEUTICALS GMBH	37
NOVARTIS AG	33
F HOFFMANN LA ROCHE AG	32
MODERNA THERAPEUTICS INC	30
TRANSLATE BIO INC	30
UNIV OF PENNSYLVANIA	26
CNRS	23
GENENTECH INC	22
UNIVERSITAT MAINZ	21
QIAGEN GMBH	19
UNIVERSITAT MAINZ	18
UNIVERSITY OF BRITISH COLUMBIA	17
COLUMBIA UNIVERSITY	17
ALNYLAM PHARMACEUTICALS INC	16
LIFE TECH CORP	16
BOEHRINGER INGELHEIM INT GMBH	16
REGENERON PHARMACEUTICALS INC	15
NOVO NORDISK AS	15
CORIXA CORP	15
HARVARD COLLEGE	14
MAX PLANCK GESELLSCHAFT	13
ISIS PHARMACEUTICALS INC	13
BRISTOL MYERS SQUIBB CO	13
UNIVERSITY OF TEXAS	13
SIRNA THERAPEUTICS INC	13
DYNAVAX TECH CORP	12
PROTIVA BIOTHERAPEUTICS INC	12
MERCK PATENT GMBH	12
IONIS PHARMACEUTICALS INC	12
GANYMED PHARMACEUTICALS GMBH	11
ASTELLAS PHARMA INC	11
THE GENERAL HOSPITAL CORP	11
BIONTECH SE	11
MASSACHUSETTS INSTITUTE OF TECH	11
ACUITAS THERAPEUTICS INC	11
MERCK & CO INC	11
MERCK SHARP & DOHME CORP	11
CHIRON CORP	10

INSERM	10
STANFORD UNIVERSITY	10

Table 3: mRNA patent landscape

A.2. Quantum

Name	**Amount of docs**
NORTHROP GRUMMAN SYSTEMS CORP	47
GOOGLE LLC	39
D WAVE SYSTEMS INC	32
IBM CORP	31
MICROSOFT TECH LICENSING LLC	28
INTEL CORP	27
QUALCOMM INC	26
TELEFON AB LM ERICSSON PUBL	24
RIGETTI & CO INC	23
FUJITSU LTD	21
SEMICONDUCTOR ENERGY LAB CO LTD	21
HEWLETT PACKARD DEV CO LP	20
NEC CORP	18
NXP USA INC	17
NXP BV	17
SONY CORP	15
CO LTD TOSHIBA	14
HITACHI LTD	14
HUAWEI TECH CO LTD	13
CEA	13
MATSUSHITA ELECT IND CO LTD	13
ALCATEL LUCENT	12
MITSUBISHI ELECT CO LTD	11
SIEMENS AG	11
KON PHILIPS ELECT NV	11
MICRON TECH INC	11
PHILIPS ELECTS NV	11
BRITISH TELECOMUNICATIONS PLC	10
AT&T CORP	9
MURATA MANUFACTURING CO LTD	9
ALIBABA GROUP HOLDING LTD	8
ID QUANTIQUE SA	8
NOKIA TECH LTD	8
SHARP CO LTD	7
ADVANCED MICRO DEVICES INC	7
NL ORGANISATIE ...	7
THOMSON CSF	7
INFINEON TECH AG	7
YALE UNIVERSITY	7
OXFORD UNIVERSITY INNOVATION LTD	7
QINETIQ LTD	6
ROBERT BOSCH GMBH	6
THE UNIV OF MELBOURNE	6

NOKIA CORP	6
LUCENT TECH INC	6
HONEYWELL INT INC	6
SAMSUNG ELECT CO LTD	6
MOTOROLA INC	6
THALES	6

Table 4: Rigetti competitors

Organization
IBM
NORTHROP GRUMMAN COR
D WAVE SYSTEMS
MASS INST OF TECH MI
BANK OF AMERICA
TOSHIBA CORP
MICROSOFT CORP
US GOVERNMENT
INTEL CORP
LOCKHEED MARTIN CORP
ALPHABET INC
RIGETTI & CO INC
SEAGATE TECH PUBLIC
NEWSOUTH INNOVATIONS
RAYTHEON TECH CORP
HONEYWELL INT INC.
MICHIGAN STATE UNIV
HP ENTERPRISE
HITACHI LTD
QUBITEKK INC
SAMSUNG ELECTRONICS
PSIQUANTUM CORP
SEEQC INC
COLUMBIA UNIV
EQUAL 1LABS INC.
YALE UNIV
RAVENBRICK LLC.
NEC CORP
STANFORD UNIV
BRIDGELUX INC.
EDICO GENOME CORP
NOKIA CORP
WELLS FARGO BANK NAT
GOVERNMENT OF ABU DH
HARVARD COLLEGE
1QB INFORMATION TECH.
QUANTUM MACHINES.
MAGIQ TECH INC
MITSUBISHI ELECTRIC
OLD AMERICAN INC
STMICROELECTRONICS.

Organization
ZYOMED HOLDINGS INC.
EVOQ NANO INC
HONDA MOTOR CO
UNIV ILLINOIS
LIGHTMATTER INC

Table 5: Patent Landscape Results of Aboy et al. (2022)

Matches: IBM, NORTHROP GRUMMAN, D WAVE SYSTEMS, MICROSOFT, INTEL, RIGETTI and CO, HITACHI, HONEYWELL INT, SAMSUNG ELECTRONICS, YALE UNIVERSITY, NEC, NOKIA, MITSUBISHI ELECTRIC.

Number of overlapping companies: 13