Cheminformatics: A Patentometric Analysis

Amit Kumar Tiwari[a,b*], Dipika Jaspal[c], Shradha Deshmukh[c], Preeti Mulay[c]

[a] *Symbiosis Centre for Research and Innovation, Symbiosis International (Deemed University), Pune, India.*

[b] *R.K. Dewan & Co., Pune, Maharashtra, India.*

[c] *Symbiosis Institute of Technology, Symbiosis International (Deemed University), Pune, India.*

meettiwari@gmail.com

Keywords: *Artificial Intelligence, Cheminformatics, Quantum Clustering, Machine Learning, Patent Analysis.*

Cheminformatics has entrenched itself as a core discipline within chemistry, biology, and allied sciences, more particularly in the field of Drug Design Discovery and Development. The article begins with a patent analysis of the progressing field of cheminformatics from 1996 to early 2021 using the Relecura and Lens patent database. It proceeds with a description of patents in various domains and aspects. The eye-catching mind map shows the landscape of cheminformatics patent search. The results reveal the star rating-wise patent counts and the trends in the sub-technological research areas. At the end of the article, quantum clustering and eminent directions towards the future of cheminformatics have been discussed. This study would provide the directions to academicians, techno enthusiasts, researchers, stakeholders, or investors and helps increase the awareness of the potential of cheminformatics and quantum clustering.

Introduction

Patents are known to be a massive single source to house technological information on product inventions and methods of inventions. Scientific articles are acknowledged and considered to be a legitimate reflection of scientific research since the beginning of time. Similarly, patents are acknowledged to reflect technological advancements and achievements. Analysis of patents related to international technical activities provides insights into technological strategies and policies implemented by various nations [1].

The organization should deep dive and understand the importance of upcoming innovation from a scientific perspective so that they can plan and invest in research and development (R&D) accordingly. These will turn help researchers in the understanding horizon of the research field. Information and knowledge regarding the field i.e. growth statistics, countries leading the development, the most prolific and impactful institutions, etc., is what decides the global technological trends and the eventual future of a field of research and innovation. Numerous fields garner researchers' attention, some of the fields are in the domain of chemistry i.e. cheminformatics.

Cheminformatics involves using computers and informational techniques and applying them in the field of Chemistry and its related issues, Drug Design Discovery, and Development. Cheminformatics is also known as interface science as it involves combining various fields such as biology, physics, mathematics, statistics, chemistry, informatics, and biochemistry. F.K Brown coined the term Cheminformatics in 1998 [2]. Cheminformatics with a combination of information channels helps in converting data into useful information that facilitates making informed decisions. Ever since drug lead identification and optimization have evolved, both "Cheminformatics" and "Chemoinformatics" have been used. but "Cheminformatics" became a popular term and gained acceptance when it was legitimized by the European Academia settled in 2006. They established in 2009, the Journal of Cheminformatics is a strong push towards the shorter variant. To understand trends in cheminformatics, the study of the patents is an important activity that can be carried out for a specific amount of time, it will help in streamlining the process and help in strategizing the road map.

Therefore, this manuscript focuses on analyzing patents statistics of 'Cheminformatics OR Chemoinformatics'. The manuscript covers patenting activity with the highest citation count, active authors, leading countries, inventors, sub-technology classifications, and recent trends in cheminformatics using Clustering along with the patent details of innovation trends and assignees. Cheminformatics is moving towards a large volume of data that requires resources and efforts beyond classical techniques [2]. For this reason, cheminformatics can take advantage of quantum clustering. Quantum clustering is an amalgamation of quantum computing and classical clustering which achieves exponential speed to process [3]. The proposed quantum algorithm for cheminformatics is discussed in the discussion section of this article.

Patenting Activity

Intellectual property (IP) has a powerful economic impact and patent law helps in governing it [4]. IP is the protection of the creation of the mind which has both moral and commercial value. This means that IP provides security for the creation of any product or knowledge-related invention. One of the types of IP is Patent. Patents play the most crucial role in legally protecting inventions introduced, researched, and developed by individuals, institutions or firms etc. The patent indicator helps in relaying information related to inventive activities, therefore, helps with patent statistics

[4]. The patent is the intangible form of IP. The owner of IP can transfer entity rights, so a rage entity that transfers the right is called the assignor, and the entity to whom the right has been transferred is called the assignee. An employer can file a patent application as an assignee if an employee gets an invention during the duration of his employment. The applicant is the one having all the legal rights on the patent application, known as the assignee. Individual, University, Organization, etc can be classified as assignee. A person who conceives an idea and converts it into an invention is called an inventor. The invention clears the below aspects [27]:

- Patentability subject matter
- Novelty
- Industrial applicability
- Inventive steps
- Specification (Disclosure of invention)

A person or firm or company who is an investor for the development of an invention cannot be considered an inventor. As they do not play any role in conceiving the invention they can be an applicant in the form of an assignee. The prior art search goal was to discover whether a certain invention is new and original so that enthusiasts could attack a corresponding patent based on its lack of novelty or lack of innovative step or based on sheer obviousness. The key is to focus on the key concepts of an invention and include the exact keywords and the synonyms of the keywords from the patent claim. There are many free patent databases available in the market such as Lens, Google patents, WIPO etc. 'Lens' is an easy-to-use patent database [5]. The search result in the Lens database has the filter option to exclude and include various attributes, or export up to 1000 patent records (in any file format). AND/OR Boolean operators can be used between targeted fields for structured search. The figure (**Figure 1**) shows various search queries using Boolean operators to fetch specific search results. The other useful database is Relecura which processes different data types, including patent data, scientific data, business data, and derives options for in-depth analysis. Relecura supports patent search and analytics, IP commercialization, identifying emerging trends, diligence for merger and acquisition, and competitive and white space analysis. Search in Relecura search was executed to import documents for further analysis, and the results were utilized to understand the growing and declining trends in the area. Finally, a competitive landscape was obtained by exporting the result.

Patent Analysis has demonstrated itself as a distinct management tool for addressing the calculated and deliberate management of the organization's technology, product, and service development process [6]. Rendering patent data into competitive analysis allows companies and firms to understand their current technical competitiveness, forecast technological trends, and swings and plan for potential competition basis innovative technologies.

A Patent system is applicable [23,24]:

- Where costs for Research and Development are high for the industries but costs of imitation are reasonably cheap.
- When competitors are not provided with suitable opportunities for innovation due to the information vouchsafed by the patents.
- When cumulative innovation requires accessing multifold snippets of knowledge managed and administered by other agents and exchange of technology becomes arduous.
- The patent needs to be licensed to a large firm as the invention developed by a small independent inventor requires considerable costs.
- Because having a secured patent is substantially more effective than having copyright. Copyright protects an individual's version or an idea but it won't protect against the independent redevelopment of that idea.

The patent can be classified as a government grant that will help the inventor to channel the use of their invention for given amount of time. It will also help the inventor in prohibiting organizations or another inventors from selling or producing inventions. Every innovation or invention must satisfy the following requirements to be patentable:

- The subject matter should be patentable.
- The invention should be useful, viz. it must have some identifiable benefits for practical applicability.
- The patent specification must contain a detailed written description of the invention, how utilized and the process of making i.e. manufacturing it in full, precise, and concise terms.
- Invention or Innovation must be not be related to nature but individual's contribution.

The patentable subject matter can be categorized as the change in any process, machine or improvement in the matter that will have an economic impact. This law helps in abiding by intellectual property.

Need for patent analysis:

- Identifies an area of research strength, weakness, and gaps thereof.
- To explore past, present, and as well as forecast future publishing trends.
- To study the productivity of the institutions/individual and the disciplines.
- Demonstration of the importance and impact of the individual interest in a research field.
- Recognizing/searching the expert or researcher in the targeted subject area.
- To identify which research is influencing its field.

Research Design

The task of accomplishing the main objective starts at the planning stage. This involves identifying information sources i.e patent database sources. Both Relecura (https://explorer.relecura.com/) and Lens (https://www.lens.org/) databases have been used in the study as they have a vast collection of patent databases that is online along with providing information related to the full-text patent collection [23,24]. Specific search criteria can be used for collection purposes (**Figure 1** - *The diagram shows the query search related to cheminformatics. Data fetch from Lens patent database (20 March 2021)* **) (All Figures are presented on Supplementary Information).

Initially, the basic "Cheminformatics OR Chemoinformatics" query has been fired with patent search in Lens research database, and 1805 Patents were found. To see the trends we have used a variety of secondary keywords i.e. Virtual Screening, Structure-Activity Relation, Machine Learning, Quantum Machine Learning, Protein Structure, High throughput Screening, Molecule, Cluster analysis, Small Molecule Libraries, drug screening, chemical structure, Quantum Clustering.

Appropriate query definition is extremely crucial because cheminformatics collects many synonyms terms. The following table shows that most of the researchers use the term "Cheminformatics" (**Table 1**). For a more precise and specific research study "Cheminformatics OR Chemoinformatics" query was used.

Table 1. Basic keyword search of Cheminformatics. Data fetched from Lens patent database (20 March 2021)

Keyword	Publication Count
Cheminformatics	1173
Chemoinformatics	752

Patents are studied by patent families as they provide good comparisons across countries for patent indicators [25]. Collection of patent applications together can be called a patent family that contains data of protection invention across various patents. This data can be represented through an analytics package or Relecura. Besides, to identify the main knowledge areas and depth of the cheminformatics, analysis on clustering can be drawn by use of RawGraphs software and illustrates a mind map using Xmind software. The mind map illustrated the overview of the cheminformatics which is covered in the result section. It also highlighted the fields of innovation in cheminformatics (**Figure 2** - *The Mind map shows the overview of cheminformatics using Lens and Relecura patent database (Mind map using the XMind software, version: XMind 8 Update 8 (R3.7.8.201807240049), access on https://www.xmind.net/) [7-15]* **). The nodes represent the Patent databases, such as Relecura and Lens, and secondary keywords used for patent analysis. Further, it shows the year-wise publications, classification, jurisdictions, and types of patent document details. **Figure 2** illustrated the trend of patent active assets. The patent assets activeness demonstrate that the assignees are interested to enforce the same, as they are mature patents. As compared to the recent patent or patent applications, they have more value [27]. Since many of the patents are granted and enforced in the current year, the patent count decreased as the year progresses. Therefore the overall patent assets also decreased. The year 2021 revealed more patent counts than the year 2040. To maintain a patent, assignees need to shell out more money as time progresses, hence the same case was with the current patent counts.

Results

Monitoring performance of the patents

For analyzing patent performance, priority date was considered as that can be considered as the closest date of invention. **Figure 3** below shows us that from 2018 onwards the inventions have been booming in the field of cheminformatics with 2020 being the year to have published and granted the highest number of patents (171 and 91 respectively).

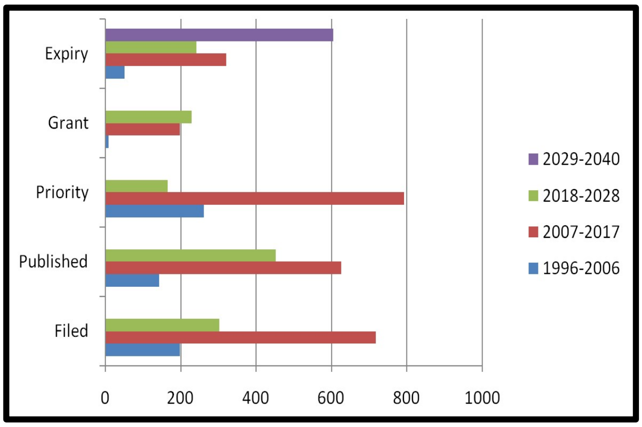

Figure 3. The bar graph shows category wise such as Filed, Published, priority, grant, expiry patent per count. Data fetch from Relecura (on 27 Mar 2021)

The country in which the inventor is based should compile patent statistics that reflect inventive activity as it is advantageous to analyze various companies' market allocation strategies [26]. The country holding the most number records is the United States of America with 598, followed by the European Patent Office (98), Japan (53), and India (52) (**Figure 4**).

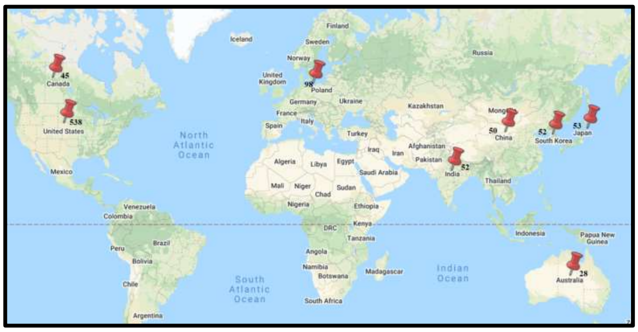

Figure 4. The world map shows the top patent filing countries in the world. Data is fetched from the Relecura patent database. (On 27 Mar 2021)

Corporate patent owners are typically assignees because the inventor assigns the invention to the company. The method and technique that is used for searching inventors are using the query (**Figure 1**). According to the analysis the Statistical Analysis System (SAS) institute from the USA had the most number of holders for cheminformatics (91 followed by Pharnext biopharmaceutical company (76), Danish multinational pharmaceutical company Novo Nordisk As (48) and USA University i.e. University of California (42). (**Figure 4**)

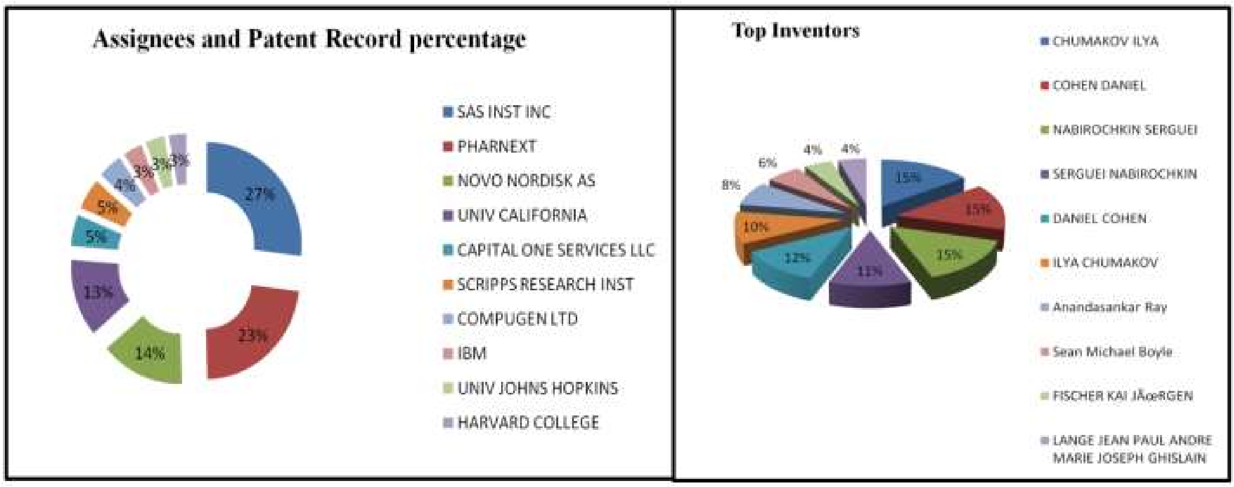

Figure 5. The chart illustrates the top assignees and inventors with their respective published patent percentages. Data fetched from Relecura patent database (On 27 Mar 2021)

Patent Classification

The system where patent examiners or groups of people code documents and publish patent applications can be called Patent classification [8]. This system makes it possible to quickly locate a document based on the technical features of the contents of the documents disclosing earlier disclosures identical. The biggest advantage of patent classification is that it enables searching documents written in various languages by use of symbols or classification syntax indifferent to words [8]. The patent classification wasn't always fixed under an agreement among people, institutions, and countries as it is today. This system was originally developed to sort paper documents but nowadays is used to search the patent database.

Each country has its benchmark and as per benchmark has a classification system set up to give some examples United States Patent and Trademark Office (USPTO) is set up by the United States, German patent classification (DPK) was set up by the German patent office/Deutsche patent in October 2010. Additionally, the European patent office (EPO) launched a joint project to create Cooperative Patent Classification (CPC) to streamline the patent classification system between two offices. European classification (ECLA) was replaced by CPC in 2013 [7].

The alphanumeric code is used to represent the IPC. The numeric code is used to represent USA classification. Patent documents or patterns are categorized and placed in a bin based on the field of invention. The bins are nothing but the classification codes.

The identification of technological fields

The IPC (International Patent Classification System) is responsible for providing standard information for categorizing and evaluating inventions based on their technological uniqueness [9]. The information provided by the said IPC constitutes is for identifying the technical domain of patents. The result of the analysis is related to the core IPC that the research study takes into account, G06F, A61P, A61K, C12Q, G01N, C07K. (**Figure 6 B**) Section G is for Physics: Computing: Calculating: Counting: Electric Digital Data Processing and G06F code contain data processing equipment or digital computing methods or functions (159). Section A is for Human Necessities: Veterinary or Medical science: Hygiene and A61P (66) and A61K (57) code is for medicinal preparation or specific therapeutic activity of chemical compounds and active ingredients without chemical characterization for example, cardiac and antiphlogistics respectively. An indicator for the scope is the variety of technical classes accredited to patents, hence for the patent's value [7] (**Figure 6 -** *The plot shows the top IPC, CPC and US codes along with the counts. The data is fetched from the Relecura patent database. (On 27 Mar 2021)* **).

The highly published grant of USA classification code is shown in (**Figure 6 C**). The US class code 702/19 (25) and 702/20 (12) represents the biological class. The 435 class belongs to chemistry, i.e. molecular biology and microbiology, which contains the highest patent publication count in various sub-fields. The 514/44 class contains drug and bio-affecting compositions with 10 published patents.

The top institutions are getting interested in the Artificial Intelligence, Biochemical, Chemical similarity, Diversity Analysis, Silico, Chemical space, Drug Discovery, Bioinformatics sub-field of study. (**Figure 7)** gives a clear picture of the institution's research interest to concerning the subfield of study along with the active authors.

The author Andreas Bender from the University of Cambridge (**Figure 7 -** *The alluvial diagram illustrates top institutions along with their respective field of study, document count and authors from scholarly works. The data is fetched from the Lens patent database. (On 17 June 2021)* **) is working in the field of Artificial Intelligence and Bioinformatics with 12 documents in his bucket [21-22]. The variety of sub-field of study can be seen in scholarly works on the Lens Patent database.

Citation Chain

The citation chain involves finding a set of research resources that are linked[28]. The citation chain can be helpful whenever getting to know about a particular research or innovation area or specifically building literature review forward and backward search strategies to find a chain consisting of many resources [17]. When a patent application is filed, the applicant may cite prior art and include older patent or non-patent literature together known as 'backward citations'. Once a patent is granted it may be cited as prior art in later patents or scholarly journals or articles, this is known as 'forward citations' [7, 17]. Forward and backward citation analysis literature is an extensive subject of in-depth study in a targeted topic. (**Figure 8 -** *The matrix plot illustrates the backward and forward citations count along with the assignee's patent number. The data is fetched from the Relecura database. (On 27 Mar 2021)* **) shows forward and backward citation counts for the field of cheminformatics. 'WO2009122128A1' has the highest number of backward citations with 28 counts and 'US10905672B2' titled "Combination of baclofen, acamprosate and medium-chain triglycerides for the treatment of neurological disorders" have 20 forward citations. (**Table 2**) shows details of the highest cited patents in Cheminformatics.

Figure 9 (*Convex hull represents star rating wise patent count. Data fetch from Relecura (On 27 Mar 2021)* **) illustrates the patent count based on the star rating. The Relecura software was explored to get the proprietary star ratings, which are usually in the range of 1 to 5. These ratings are based on a composite metric, and combinations of technologies, business value, litigation etc related factors. The majority of the patent documents have low ratings, only a few documents, that have more techno commercial prospects have the higher rating. The highest star ratings are for 23 documents, which comes under the category 3.5 to 4.5 collectively. Most of the patent documents (355) have a 1.5 rating, while 17 patent documents had lowest star rating i.e. 0.5 and appear to be insignificant from a valuation point of you.

Table 2. Details of highly cited patents. The data is fetched from the Lens patent database (15 March 2021)

Sr. No	**Patent Document**	**Applicants**	**Inventors**	**Published Date**	**Citation**	**Patent Application No.**	**Lens ID**
1	"System And Method For Knowledge Retrieval, Management, Delivery And Presentation"[12]	Omoigui Nosa	"Omoigui Nosa"	18-Mar-10	1135	US2010/0070448 A1	046-419-976-473-363
2	"Methods And Systems For Annotating Biomolecular Sequences"[13]	Compugen Ltd	"Mintz Liat, Xie Hanqing, Dahari Dvir, Levanon Erez, Freilich Shiri, Beck Nili, Zhu Wei-yong, Wasserman Alon, Hermesh Chen, Azar Idit, Bernstein Jeanne"	12-Apr-07	317	US2007/ 0083334 A1	198-811-424-935-775
3	"System And Method For Providing Educational Related Social/geo/promo Link Promotional Data Sets For End-User Display Of Interactive Ad Links, Promotions And Sale Of Products, Goods, And/or Services Integrated With 3d Spatial Geomapping, Company And Local Information For Selected Worldwide Locations And Social Networking"[14]	Heath Stephan	Heath Stephan	21-Mar-13	274	US2013/0073387 A1	117-540-671-976-929
4	"Human Thrombospondin Polypeptide"[15]	Compugen Ltd	"Mintz Liat, Xie Hanqing, Dahary Dvir, Levanon Erez, Freilich Shiri, Beck Nili, Zhu Wei-yong, Wasserman Alon, Bernstein Jeanne"	29-Jun-10	266	US7745391 B2	024-164-474-660-548
5	"Data Pool Architecture, System, And Method For Intelligent Object Data In Heterogeneous Data Environments"[16]	Biosentients Inc	"Stanley Robert A, Gombocz Erich A"	01-Jan-04	239	US2004/0003132 A1	019-441-425-316-179

Discussion

Clustering Methods

One of the most useful and innovative ways to learn incrementally about the ever-growing data is to cluster the data using data-clustering methods / techniques [29]. Due to the advent of technology, a huge amount of data is getting generated which requires analysis, predictions and forecasting. Clustering will primarily form groups of similar data items based on how close they are as far as their impactful features are concerned. The basic groups or clusters are formed in initial phases and then on influx of new data existing clusters can be updated / appended or new clusters are formed [30]. The learning about the data starts from the initial cluster formation itself and continues as the clusters are either appended or new clusters are formed. Every cluster has its own characteristics / features including center of the cluster, cluster range, threshold, closeness, distance between cluster members, distance among clusters etc to name a few. Where there is a huge collection of ever-growing dynamic data, the learning is utmost essential and incremental clustering will be the best solution. And hence data clustering can be applied to data from any domain.

The clustering problem is defined as, given data I = {I1, I2,…., In} and k is the integer value, so the clustering is to perform mapping of Y: I → {1,2,…,k}, where each point Ix, x Е {1,2…n} where Cj is the cluster and point assigned to the Cj, j Е {1,2….,k}. A cluster Cj mapped to Cj = {Ix |Y(Ix) = Cj, Ix Е I}.

There are varieties of options available to sort N objects into K groups. The role of the clustering method is to find a suitable set of club classes that shows the unseen structure in the data. However, different clustering methods are available, and produce different cluster shapes because they are differing of empirical and theoretical base. Evaluation of the clustering methods depends on the following measures of clustering, which are to be considered while creating categories of objects with the aim of data storage, retrieval, analysis, learning, predictions and others [29-31]:

- Choosing a relevant clustering method
- Selecting suitable similarity measure from available clustering methods
- Requirement update for dynamic data must be considered
- While retrieving the clusters, the cluster hierarchy must be selected
- Evaluate the effectiveness of the obtained results

The types of requirements of clustering in complex large datasets are [31]:

- High dimensionality: Handle high dimensional data effectively
- Shape : Control any shape of the data
- Noisy data: Insensitive to outliers
- Scalability: Linear to the data

Cheminformatics and Clustering

Quantum technology will bring the existing fundamental base technology and integrate it with the existing engineering and learning technology. Quantum computing goes to the subatomic level which completely transforms the way computers operate. Quantum computers that are used in practice can solve extremely complex problems [3]. Cheminformatics is a very useful and highly contributing field. Cheminformatics is extremely helpful in designing and discovering drugs to avoid the expensive trial and error method. This field also made it easy for scientists to search for molecules, the molecular structure, and the molecule's detailed information across a wide range of databases. The cheminformatics experts in the pharmaceutical industry are facing challenges regarding the analysis of a large collection of molecules. The power of quantum technology can help to analyze complex sets of molecules within a small amount of time [18-19]. The engagement of clustering and quantum technology is needed to analyze molecules for compound purchasing, virtual screening, or processing of high-throughput screening results. Until a few years ago, data sets comprised of a few 100 items. Nowadays technologies are capable of storing and processing a much larger amount of data. This makes it difficult and at times impossible for traditional data processing units to process these datasets. Datasets eventually become humongous due to their volume, velocity, and variety. It is beyond the capabilities of the existing IT systems as they aren't robust enough to analyze, process, and store these databases. Cheminformatics is a scientific branch where machine learning is applied to solve (Computational methods) numerous chemistry problems [10]. Cheminformatics comprises of extremely challenging problems like the few mentioned below such as locating important molecule features from the 3D molecular structures with the help of an ML model for better and accurate predictions [10]. This is where quantum clustering enters the picture. As we know up until a few years ago most of the research work in these fields was largely theoretical, we now have demonstrable Quantum Machine Learning algorithms at our disposal. Quantum computers can have an extremely vast and lasting effect on machine learning.

Quantum Machine Learning is considered an approach that will lead to innovation and a great future because it solves unique problems, the need to constantly process huge amounts of data as that's the society we live in where humongous data collection and processing happens invariably and where novel research methods can have a tremendous effect on both economy and life. Quantum clustering is an inversion problem of quantum mechanics [20].

The Quantum Incremental Clustering (QIC) method has many benefits. It possesses the unique ability to recognize clusters of different shapes and subsequently determines their center. It doesn't necessarily require prior knowledge regarding the cluster members to do this. The QIC algorithm applies to complex datasets as well as to high-dimensional datasets.

Clustering in Cheminformatics

Figure 10 (*The linear dendrogram shows the top institutions' file patents along with the patent count, top publisher, active authors, patent citation, scholarly citation respectively. The data is fetched from the Lens patent database (17 March 2021)* **) below shows top institutions that are working on the "cheminformatics and clustering" research area and have published patents. The University of Cambridge has published 10 patents which received 763 scholarly citations and 3 patent citations. The University of Manchester has also published 10 patents, their top author is Douglas B. Kell and their top publisher is Frontiers Media S.A. They have received 1 patent citation and 298 scholarly citations so far. Although the European Bioinformatics Institute has published (6 patents) a patent less than that of the University of Sheffield (7 patents), it has more Scholarly citations (205 citations) than that of the University of Sheffield (159 citations). BioMed Central and Steinbeck is their top publisher and active author respectively. Meanwhile, the University of Sheffield has Wiley as their top publisher and Peter Willet as their top author. Katholieke University Leuven also has 6 published patents but only 190 patent citations. Springer Nature is their top publisher and Jan Ramone is their active author. The University of Sheffield, the European Bioinformatics Institute and the Katholieke University Leuven do not have any patent citations.

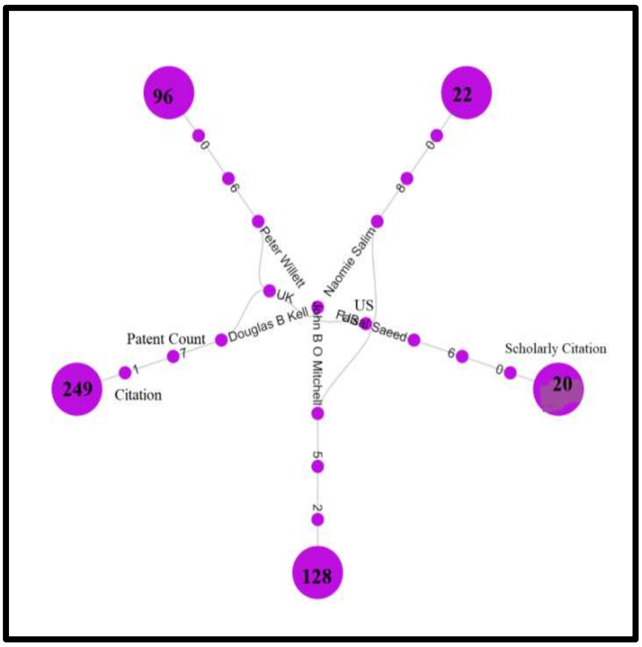

Figure 11. The circular dendrogram shows the top active authors with country, Patent count, patent citation, and Scholarly citations respectively. Data fetch from Lens patent database (14 March 2021)

Figure 11 shows that Douglas B. Kell has the highest number of scholarly citations with 249. He has 7 published patents with 1 patent citation. Meanwhile, Peter Willet has 96 scholarly citations and 6 published patents. Together they help the UK contribute 345 scholarly citations. While on the other hand, the US has a total of 170 scholarly citations. John B O Mitchell contributes the most to this with 128 scholar citations, 2 published patents, and 5 patent citations. Naomei Salim has 22 scholarly citations and 8 published patents on her name and Futsal Saeed has 20 scholarly citations and 6 published patents, with uncited patents.

Therefore, cheminformatics deserves more attention and in-depth study for paramount future discoveries[21-22].

- As science progresses the number of cheminformatics tools is increasing, bypassing the old tools and resources, creating more innovation opportunities for a better world.
- Chemical features could be crucial for the performance of any drug or sequence labeling tasks so researchers would explore numerous structural, categorical, or contextual features to enrich the classical as well as quantum cheminformatics clustering model.
- The end-stage molecular profiles for each disease can be clustered from which new insights can be evaluated to provide earlier diagnosis and treatments to the patients.
- The structural cluster similarity information strengthens the prediction of the base kernel when the dataset is consequently structurally diverse and sparse. The performance of the structural cluster kernel can be enhanced or boosted by considering a large number of unlabeled instances.
- The classical clustering method forms structure in low-dimensional space but the QC method forms structure in high dimensional feature space. The QC method is impactful to extend the study of various types of solar cells for further analysis of photovoltaic space and other regions of materials space.

The set of data points in a particular feature space can be related to the Schrodinger equation for which its potential can be identified by data. This can be one example of a clustering solution.

Here is proposed a new approach to the QIC scheme using Quantum tomography. The proposed QIC used classical Incremental Clustering to group data as per moving points to the nearest local minimum of the Vector point. The primary focus of QIC is to find the dynamic distances to achieve a trajectory relationship between data points. The proposed QIC algorithm for Cheminformatics is relevant to organizations and universities for R&D purposes. The approach is more in terms of looking at technology trends that provide future direction to R&D in the area of Cheminformatics. The innovative startups also refer to the relation between clustering and Cheminformatics to understand the focus areas and where startups can take up R&D.

Conclusion

Patent analysis provides certain features which are unique to organize and study while making a patent landscape in a particular area of technology. This manuscript presented the patent landscape for the emerging cheminformatics field using the two patent databases, i.e. Relecura and Lens. The output of the patent analysis discussed in this research design and result section shows the patenting activity and technological trends in a cheminformatics field. Overview of the patent landscape illustrates the need for the patent analysis to step up in the cheminformatics innovative world. The state-of-art search related to patents using the primary query "Cheminformatics OR Chemoinformatics" with the secondary query "Cheminformatics OR Chemoinformatics AND Clustering". The result section provides certain technical references of highly cited patents in the Lens patent database concerning Cheminformatics. The information that was extracted from a different patent database can serve the popular technological area and sub-technological areas under the cheminformatics field. To understand the different facets of cheminformatics patent analysis, it is crucial to know how the cheminformatics technology trends are mapped. Therefore, mapping of patents involves a multistep process, which includes the entire collection of patent documents, screening of those patent documents, organization or clusterization or classification of patent documents, and analysis of patent documents. The manuscript gives the details from the point of view of what are the innovations in cheminformatics area; USA country is in lead with SAS institute, top active authors & inventors and a glimpse of the citation chain. The discussed patent analysis of cheminformatics will be relevant to companies as well as research institutions. Also, the investors to identify the gaps between patents and those breach provides an opportunity for further research. The future direction of QIC is given to contribute to the cheminformatics innovative field. The study of patents in the field of Cheminformatics is important as far as Drug Design, Discovery and Development is concerned.

** *All Figures are presented on Supplementary Information*

References

[1] Rozhkov S, Ivantcheva L. Scientometrical indicators of national science & technology policy based on patent statistics data. World Patent Information 1998;20(3-4):161-166.

[2] Brown F. Chemoinformatics: What is it and How does it Impact Drug Discovery. Annual Reports in Medicinal Chemistry 1998;33:375-384.

[3] Wang S, Sakk E. Quantum Algorithms: Overviews, Foundations, and Speedups. 2021 IEEE 5th International Conference on Cryptography, Security and Privacy (CSP) 2021:17-21.

[4] Grimaldi M, Greco M, Cricelli L. A framework of intellectual property protection strategies and open innovation. Journal of Business Research 2021;123:156-164.

[5] Parikh M. Unleashing bioprinting technology through patent intelligence. Drug Discovery Today 2021;26(6):1547-1555.

[6] Sick N, Merigó J, Krätzig O, List J. Forty years of World Patent Information: A bibliometric overview. World Patent Information 2021;64:102011.

[7] OECD Patent Statistics Manual. OECD Publishing, Paris 2009.

[8] Iwayama M, Fujii A, Nanba H. Challenges in Patent Information Retrieval. Evaluating Information Retrieval and Access Tasks 2020;3:49-69.

[9] Sanju S, Sankaran S, Achuthan K. Energy Comparison of Blockchain Platforms for Internet of Things. 2018 IEEE International Symposium on Smart Electronic Systems (iSES) (Formerly iNiS) 2018:235–238.

[10] Pirashvili M, Steinberg L, Belchi Guillamon F, Niranjan M, Frey J, Brodzki J. Improved understanding of aqueous solubility modeling through topological data analysis. Journal of Cheminformatics 2018;10(1):54-62.

[11] Dong J, Cao D, Miao H, Liu S, Deng B, Yun Y, Wang N, Lu A, Zeng W, Chen A. ChemDes: an integrated web-based platform for molecular descriptor and fingerprint computation. Journal of Cheminformatics 2015;7(1):1-10.

[12] Omoigui N. System And Method For Knowledge Retrieval, Management, Delivery And Presentation. US2010/0070448 A1. 2010

[13] Mintz L, Xie H, Dahari D, Levanon E, Freilich S. Methods And Systems For Annotating Biomolecular Sequences. US2007/ 0083334 A1 2007

[14] Heath S. System And Method For Providing Educational Related Social/geo/promo Link Promotional Data Sets For End-User Display Of Interactive Ad Links, Promotions And Sale Of Products, Goods, And/or Services Integrated With 3d Spatial Geomapping, Company And Local Information For Selected Worldwide Locations And Social Networking. US2013/0073387 A1. 2013

[15] Mintz L, Xie H, Dahari D, Levanon E, Freilich S. Human Thrombospondin Polypeptide, US7745391 B2. 2010

[16] Stanley R, Gombocz E. Data Pool Architecture, System, And Method For Intelligent Object Data In Heterogeneous Data Environments, US2004/0003132 A1. 2004

[17] Von Wartburg I, Teichert T, Rost K. Inventive progress measured by multi-stage patent citation analysis. Research Policy 2005;34(10):1591-1607.

[18] Romero J, Babbush R, McClean J, Hempel C, Love P, Aspuru-Guzik A. Strategies for quantum computing molecular energies using the unitary coupled cluster ansatz. Quantum Science and Technology 2018;4(1):014008.

[19] Lenselink E, Ten Dijke N, Bongers B, Papadatos G, Van Vlijmen H, Kowalczyk W, IJzerman A, Van Westen G. Beyond the hype: deep neural networks outperform established methods using a ChEMBL bioactivity benchmark set. Journal of Cheminformatics 2017;9(1):1-14.

[20] Wittek P. High-performance dynamic quantum clustering on graphics processors. Journal of Computational Physics 2013;233:262-271.

[21] Wegner J, Sterling A, Guha R, Bender A, Faulon J, Hastings J, O'Boyle N, Overington J, Van Vlijmen H, Willighagen E. Cheminformatics. Communications of the ACM 2012;55(11):65-75.

[22] Mak L, Marcus D, Howlett A, Yarova G, Duchateau G, Klaffke W, Bender A, Glen R. Metrabase: a cheminformatics and bioinformatics database for small molecule transporter data analysis and (Q)SAR modeling. Journal of Cheminformatics 2015;7(1):1-15.

[23] Penfold R. Using the Lens database for staff publications. Journal of the Medical Library Association 2020;108(2):341.

[24] Frazier R, Carter-Templeton H, Wyatt T, Wu L. Current Trends in Robotics in Nursing Patents—A Glimpse Into Emerging Innovations. CIN: Computers, Informatics, Nursing 2019;37(6):290-297.

[25] Kabore F, Park W. Can patent family size and composition signal patent value?. Applied Economics 2019;51(60):6476-6496.

[26] Somaya D. How Patent Strategy Affects the Timing and Method of Patent Litigation Resolution. Advances in Strategic Management 2016:471-504.

[27] Von Wartburg I, Teichert T, Rost K. Inventive progress measured by multi-stage patent citation analysis. Research Policy 2005;34(10):1591-1607.

[28] Nomaler Ö, Verspagen B. Knowledge flows, patent citations and the impact of science on technology. Economic Systems Research 2008;20(4):339-366.

[29] Nalinipriya G, Geetha M, Cristin R, Maram B. Biomedical data mining for improved clinical diagnosis. Artificial Intelligence in Data Mining 2021:155-176.

[30] Barrera-Vázquez O, Gómez-Verjan J, Magos-Guerrero G. Chemoinformatic Screening for the Selection of Potential Senolytic Compounds from Natural Products. Biomolecules 2021;11(3):467.

[31] Thrun M, Stier Q. Fundamental clustering algorithms suite. SoftwareX 2021;13:100642.