excelra

Where data means more

Patents are the Cornerstone for Advancing Data-Driven KRAS Drug Discovery

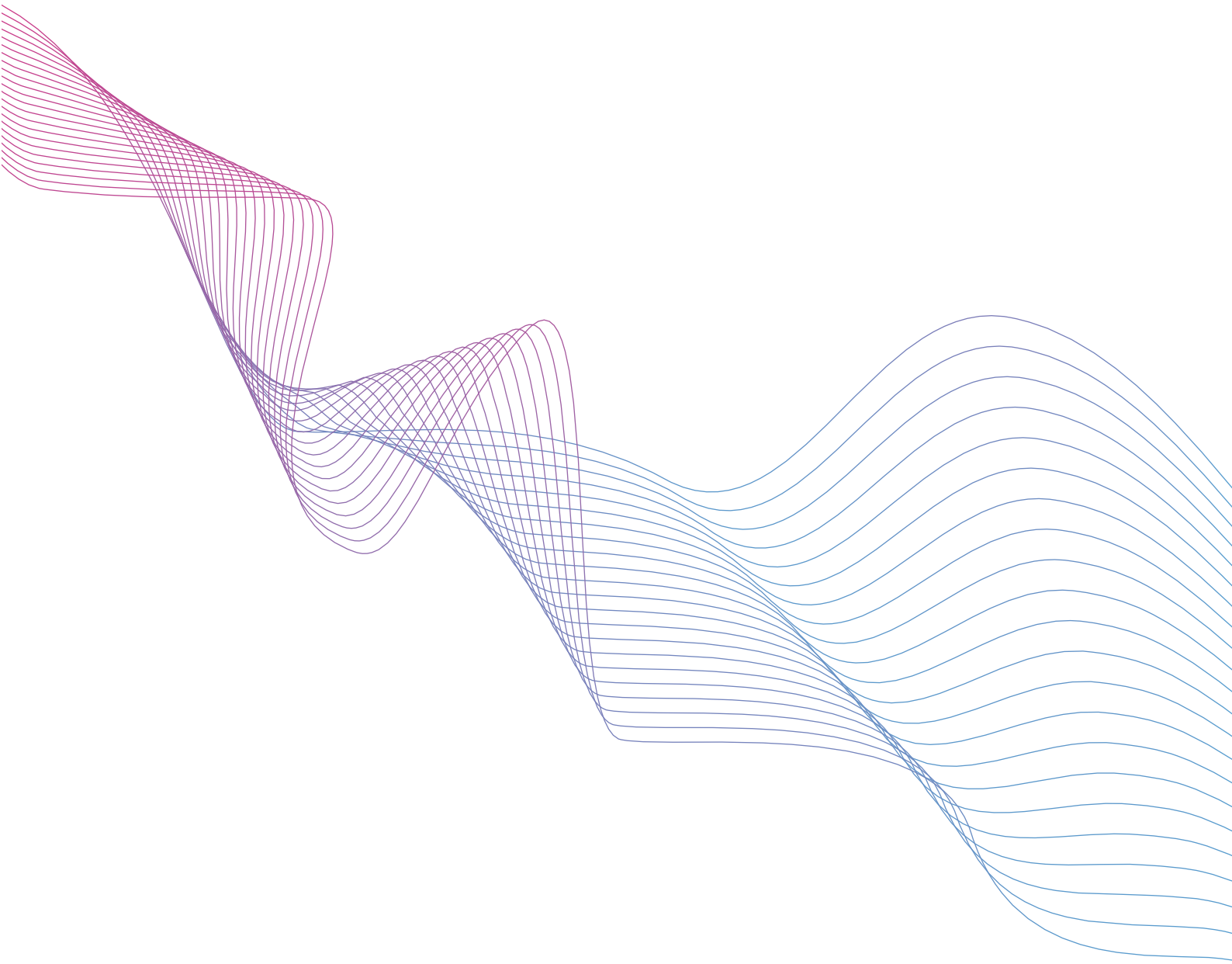
Supporting the KRAS research with GOSTAR® data & custom patent curation services-a case study

CASE STUDY

Introduction

The Kirsten rat sarcoma viral oncogene homolog (KRAS) is a member of the RAS oncogene family, which encodes a protein called K-Ras. The protein functions as a key component of the rat sarcoma virus/mitogen-activated protein kinase (RAS/MAPK) pathway, which transmits extracellular signals to the nucleus, and regulates cell growth, proliferation, and differentiation[1]. Dysregulated KRAS has been linked to cancer progression, including tumor growth, cancer cell survival, invasion, and migration[2]. KRAS mutations are among the most common genetic alterations found in human cancers, often leading to resistance against targeted therapies and unfavorable treatment outcomes[3].

Historically, targeting KRAS mutations has presented significant challenges. However, recent advances, exemplified by the development of Sotorasib, have led to a shift in treatment options. The FDA approval of Sotorasib for the treatment of KRAS-G12C mutant non-small cell lung cancer (NSCLC) represents a major advance in cancer therapy[4]. Ongoing efforts persist in the development of effective small-molecule KRAS inhibitors, aimed at overcoming past setbacks and utilizing novel binding sites.



Client's challenge

Our client, a leading US biopharmaceutical company, is actively involved in KRAS research. A key aspect of their R&D involves gathering novel chemical structures and associated bioactivity data from patents related to KRAS to fuel their in-house data-driven discovery programs. Patents are often considered a rich source of diverse sets of chemotypes, reflecting the broad scope of industrial innovation, while journal articles tend to focus on specific research questions, resulting in a narrower set of compounds. The number of patent applications featuring the term KRAS in either title, abstract, or claim in classes A61 and C07 grew from 2,743 in 2014 to 74,983 in 2021, indicating a significant rise in interest in KRAS-modulating small molecules. A meta-analysis of patents filed and granted between 2014-2023 revealed a total of 943,512 patents across various patent classes.

When comparing databases, GOSTAR®, renowned for its patent-derived compounds, offers 24 times as many unique compounds as ChEMBL, which primarily extracts data from journal articles[5]. This underscores GOSTAR®'s position as an industry leader in the extraction and structuring of structure-activity relationship (SAR) data from patents.

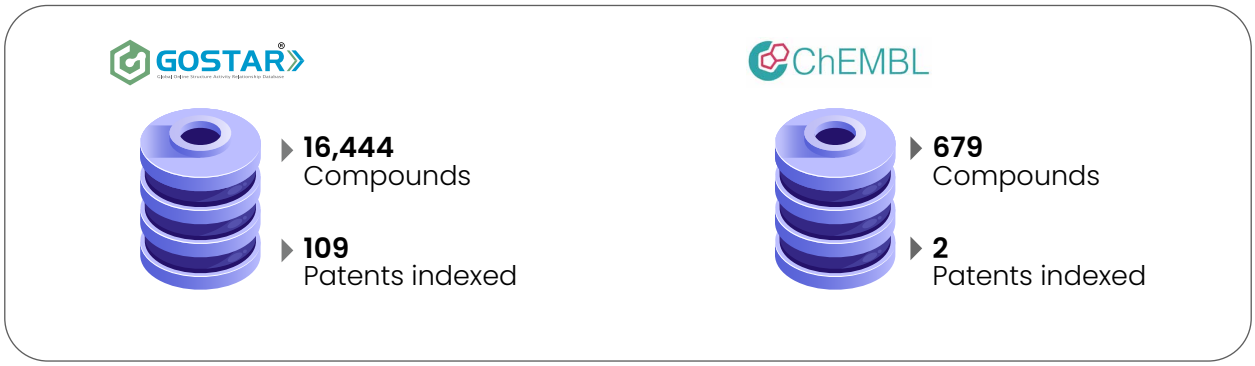


Figure 1: GOSTAR® vs. ChEMBL stats for Human KRAS

To visualize the chemical space diversity between these datasets, we utilized ChemTreeMap, an interactive tool for exploring chemical space. This tool combines extended connectivity fingerprints and a neighbor-joining algorithm to generate a hierarchical tree with branch lengths proportional to molecular similarity, akin to phylogenetic trees commonly used in biology[6].

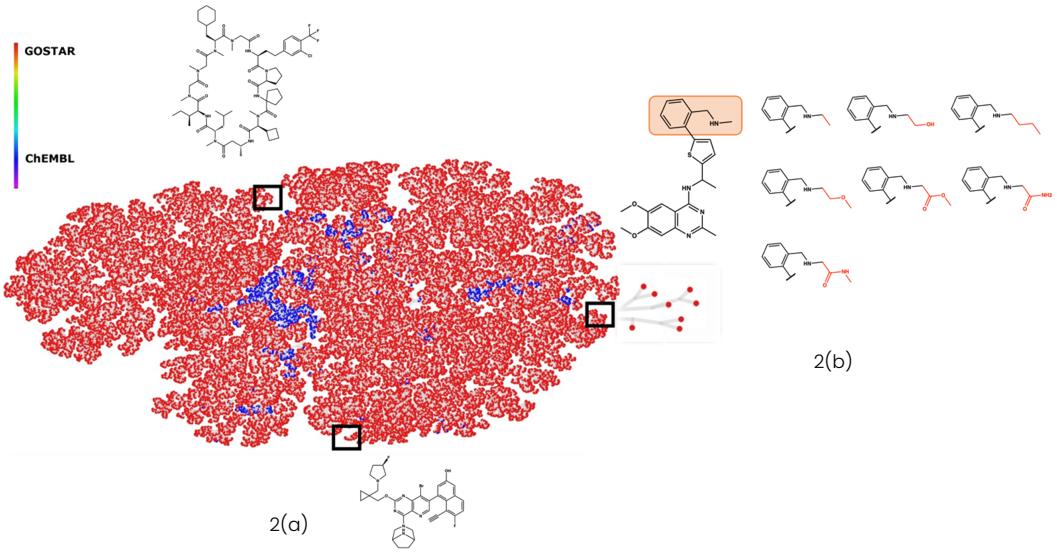


Figure 2: (a) Chemical diversity visualization - GOSTAR® vs. ChEMBL
(b) Example tree illustrating a cluster of closely related structural analogs.

The figure above depicts a treemap illustrating the diversity of chemical space. Red dots represent compounds from the GOSTAR® database, while blue dots represent those from ChEMBL. Each branch in the treemap represents a cluster of similar compounds. The visualization reveals that the GOSTAR® chemical space is more expansive and diverse, while ChEMBL covers a more limited chemical space with a smaller dataset, which reinforces the fact that patents hold diverse compound sets when compared to journals.

Patents play a crucial role as an information source for informed decision-making in drug discovery. Researchers must maintain an up-to-date internal repository of data sourced from these patents, enabling them not only to drive their medicinal chemistry programs effectively but also to assess the uniqueness of their solutions compared to competitors and swiftly identify potential opportunities.

Fortunately, the client was aware of our capabilities in this field. We are the global leader in the manual extraction of SAR data from scientific literature and have been providing data solutions to pharmaceutical companies for 20 years. The client engaged us to extract and validate the required content and deliver clean, consistent, analysis-ready data.

Our approach

With over 60 PhDs in our data curation team, we possess the domain expertise demanded by our clients to identify relevant literature, extract appropriate data, and deliver it in a standardized, analysis-ready format. Collaborating with Excelra will help address the challenges faced by researchers, particularly the need for manual curation, enabling them to focus on the formidable task of identifying novel compounds with KRAS-modulating profiles.

Manual extraction by researchers

15-20h
Weekly time spent extracting, curating, and structuring SAR data from patents

Impact of collaborating with Excelra

60-70%
Reduced time for data extraction and curation

9-14h
Saved time for researchers to refocus on core research activities

Figure 3: Excelra's patent curation services accelerate drug discovery programs

Our curation process includes three stages of data extraction: manual curation, review, and quality control. Thanks to our combination of scientific expertise and technical excellence, we were able to collate, prepare, and deliver the client's data set in an exceptionally short time. We ensured the manually extracted data included exemplified chemical structures and associated experiments, reported for a variety of assays.

The results

Meeting the client's requirement, we delivered a comprehensive data set that could immediately be added to the internal repository for further review and analysis. Their supplied data included:

- Chemical structure
- Compound's chemical and IUPAC name
- Biological source or Species
- Experimental endpoints with prefixes and units of measurements
- InChI
- Targets
- Cell lines
- Assay conditions

With our assistance, the client was able to swiftly proceed with the ongoing research program, avoiding the bottleneck of the data collection phase. The efficacy of our manual curation and quality control was greatly appreciated, and the client has repeatedly returned with similar data curation requests. As the global leader in the manual extraction of SAR data from scientific literature, we regularly replicate these results with some of the world's most successful life science companies. All of our clients understand the value of streamlining their data collection and standardization processes, and we are constantly updating our libraries and expanding our capabilities so more pharmas and biotechs can benefit from the time and cost savings we deliver. If you have data curation requirements, we can provide what you need.

Whatever your objectives, we'll help you achieve them.

References

1 Tsuchida, N.; Ryder, T.; Ohtsubo, E. Nucleotide Sequence of the Oncogene Encoding the p21 Transforming Protein of Kirsten Murine Sarcoma Virus. *Science* **1982**, *217* (4563), 937–939. DOI:10.1126/science.6287573

2 Bauml, J. KRAS: Kirsten RAT Sarcoma Viral Oncogenic Homolog.https://www.healio.com/hematology-oncology/learn-genomics/lung-cancer-assessment-treatment/kras-kirsten-rat-sarcoma-viral-oncogenic-homolog (accessed Feb.14, 2024)

3 Lièvre, A.; Bachet, J. B.; Le Corre, D.; Boige, V.; Landi, B.; Emile, J. F.; Côté, J. F.; Tomasic, G.; Penna, C.; Ducreux, M.; Rougier, P.; Penault-Llorca, F.; Laurent-Puig, P. KRAS Mutation Status Is Predictive of Response to Cetuximab Therapy in Colorectal Cancer. *Cancer Research* **2006,** 66 (8), 3992–3995. DOI: 10.1158/0008-5472.CAN-06-0191

4 Jaber, N. Sotorasib is First KRAS Inhibitor Approved by FDA. https://www.cancer.gov/news-events/cancer-currents-blog/2021/fda-sotorasib-lung-cancer-kras (accessed Feb.14, 2024)

5 ChEMBL Database. http://www.ebi.ac.uk/chembl/ (accessed Feb.14, 2024)

6 Lu, J.; Carlson, H. A. ChemTreeMap: an interactive map of biochemical similarity in molecular datasets. *Bioinformatics* **2016**, 32 (23), 3584–3592. DOI: 10.1093/bioinformatics/btw523