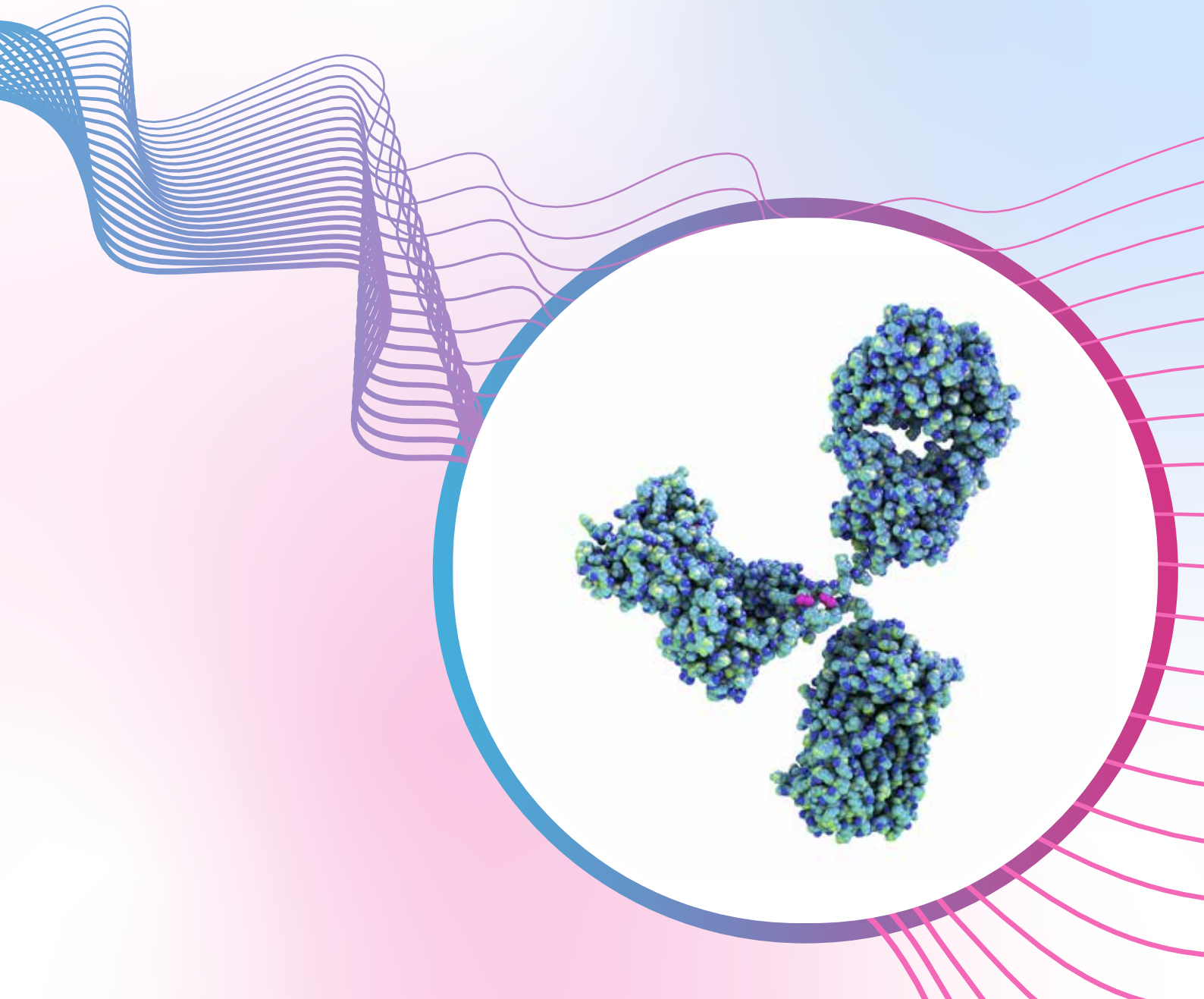
excelra

GOSTAR
Global Online Structure Activity Relationship Database

Unveiling the molecular code

Antibody sequence mining and target affinity analysis

CASE STUDY



Purpose

Comprehensive extraction of clones, sequences, and target affinity information from patents

Client

Industry
AI/ML

Location
APAC

Services used
Curation services;
Product design and development

Results

AI/ML data for precision medicine

Strategic leveraging of antibody sequence and target affinity data, structured and annotated meticulously, empowered the creation of robust training data sets. These data sets serve as the bedrock for training AI and advanced machine learning technologies, enabling the achievement of precision and personalized medicine.

Data provides valuable insights into the sequences of amino acids within antibodies, which directly influence their binding affinity to targets. This deep understanding of amino acid sequences empowered the client to unlock the full potential of antibodies and advance precision medicine.

Upon analyzing antibody affinity data, the client identified antibodies that exhibit promising interactions with multiple targets beyond their original intended use. This knowledge allowed the client to explore the repurposing of approved antibodies, utilizing their existing characteristics and expanding their therapeutic potential to address different diseases.

Specification

The client is an artificial intelligence (AI) drug discovery company bringing a magical revolution to drug discovery journey. From customized target identification to lead generation, their workflow AI platform generates the best insights for every step to develop commercially valuable drugs from in-house and partnership projects.

They have achieved remarkable success in AI drug discovery, leveraging its proprietary technology platform to accelerate the drug discovery process and develop novel treatments for complex diseases. They would like to remain committed to their mission of revolutionizing drug discovery through AI and deliver better treatments to patients worldwide.

The customer has a critical need for a comprehensive training set focused on therapeutic monoclonal antibodies (mAbs) in conjunction with structure-activity relationship (SAR) data. Their objective is to leverage this dataset to enhance their AI/ML algorithm or model, enabling the identification of new targets within the field of immune oncology. However, the current availability of such a specialized training set, specifically tailored to their requirements, presents a significant challenge.

The lack of such a dataset has profound consequences, including limited target identification, compromised AI/ML performance, missed opportunities, and delayed innovation.

Our approach

The client partnered with us due to our established legacy in data curation. We have a proven track record of bringing structure to the vast amount of data available, transforming it into valuable and actionable insights.

We follow a comprehensive and rational approach to meet our client's requirements. Here are the key components of our proposed solution:

Defining the project scope: We have clearly defined the scope of the project, outlining the specific objectives and deliverables to ensure alignment with the client's needs.

Creation of a data extraction template: To streamline the data curation process, we have developed a structured data extraction template. This template includes all mandatory data variables, ensuring consistency and completeness in the curated dataset.

Identifying data sources: We have conducted thorough research and identified relevant and reliable data sources that contain the necessary information on therapeutic monoclonal antibodies and their binding targets.

Protocol documentation: We have created detailed protocol documentation to maintain high-quality standards throughout the data curation process.

Data variable identification: We have identified the essential data variables that are crucial for the client's AI/ML algorithm/model, ensuring that the curated dataset is comprehensive and aligned with the desired outcomes.

Data delivery in excel format: Upon completion of each target curation, we have delivered the curated data in Excel format. This format ensures compatibility and ease of integration with the client's existing systems and processes.

Results

Our proposed solution focused on building a repository of therapeutic monoclonal antibodies against their binding targets. The curated dataset will encompass the following key data sections:

Clone details: This section has provided information on the specific clones, including their specificity, orientation, and the corresponding binding targets.

Sequence details: We have delved into the sequence details of the monoclonal antibodies, with a particular emphasis on the variable light (VL) and variable heavy (VH) regions. This information is vital for understanding the molecular characteristics and functional properties of the antibodies.

Binding affinity information: We have included quantitative results and methodologies related to the binding affinity of the antibodies. This data has shed light on the strength and specificity of the antibody-target interactions.

Stability parameters: Our curated dataset encompasses stability parameters such as thermostability and pharmacokinetics. These parameters provide insights into the stability and recycling of the antibodies, essential considerations in drug development.

During our curation process, we identified several major targets of interest. However, we didn't limit ourselves to these specific targets but also considered other available targets mentioned in the patents, as well as focused targets for curation. This approach provides our customer with a comprehensive understanding of the targets that can be studied or further explored. The key targets we identified include:

PD1	TIGIT	PCSK9	BAFF	EGFR
PDL1	CD3	TNF-alpha	VEGFA	HER2
CTLA4	CD52	BCMA	CD202	C5

By expanding the scope beyond named targets and incorporating additional targets from the patent data, we have enriched the dataset and opened possibilities for further analysis and investigation.

This broader perspective allowed our client to gain insights into a wider range of potential targets and make informed decisions regarding their research and development strategies

Here are the high-level curation statistics from the dataset we curated exclusively for the customer, tailored to their specific requirements:

538 Total patents curated	**24,222** Total data rows curated	**9,705** Total antibody sequences
7,289 Unique variable heavy (VH) chain sequences	**4,862** Unique variable light (VL) chain sequences	**234** Applicants/companies
71 Methods	**105** Cells/Cell lines	**267** Targets

These numbers highlight the scale and breadth of our curation efforts, resulting in a rich and diverse dataset that provided valuable insights into the field of therapeutic monoclonal antibodies.

Conclusion

The success of this project can be attributed to several key factors. Firstly, our systematic approach to data extraction, guided by a well-defined project scope, allowed us to efficiently navigate the complex landscape of patent information. By leveraging our experience in data curation, we meticulously extracted the desired antibodies and their associated sequences and target affinity data, ensuring the highest level of accuracy and completeness.

Our capabilities extend beyond antibody data and encompass the curation of oligomers, siRNAs, peptides, and antibody drug conjugates (ADCs).