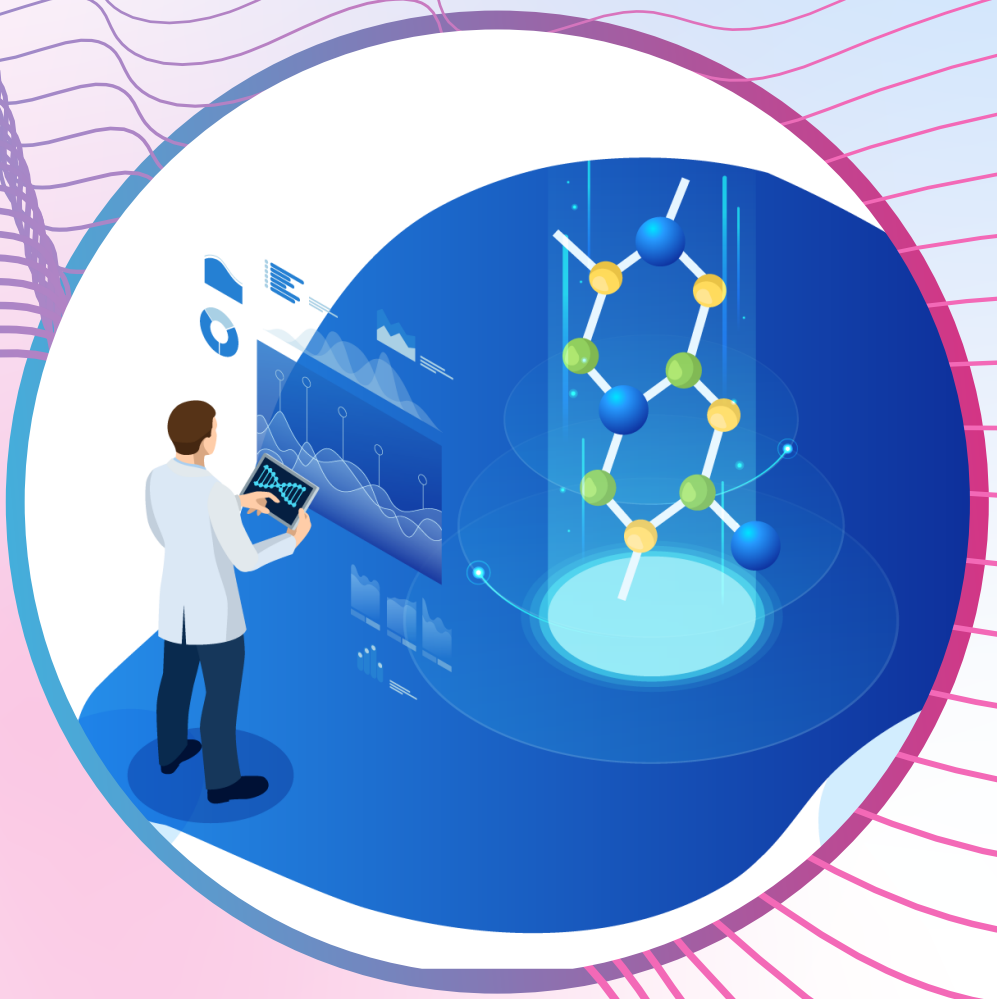
excelra

Structured and analysis ready data for AI/ML based drug discovery

CASE STUDY



Purpose

A biotech company was interested to employ AI/ML technologies to identify potential small molecules for therapeutic development in the areas of oncology and renal fibrosis.

Client

Industry
Biotech

Location
US

Sector
Multiple

Client requirement

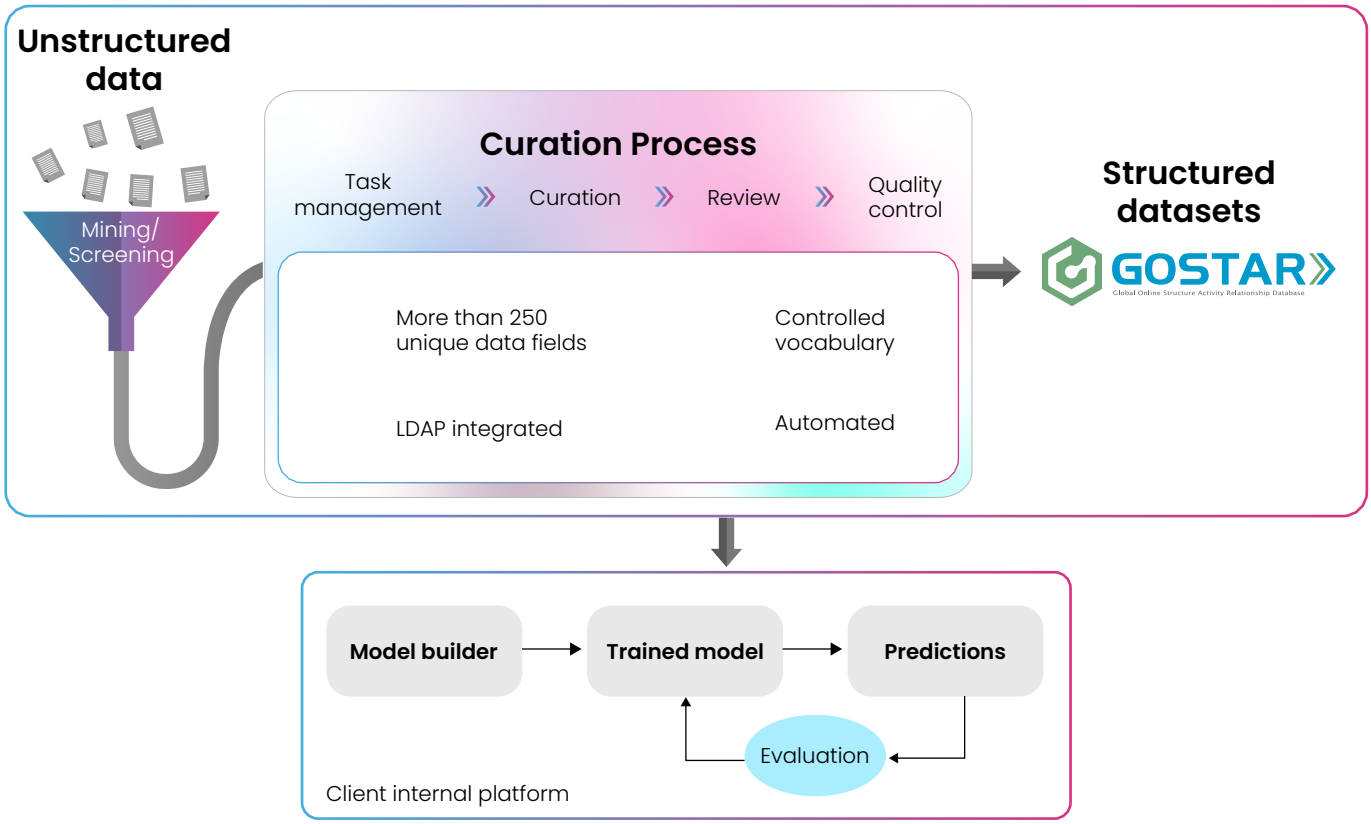
The client required high-quality, harmonized and structured datasets of small molecules, encompassing comprehensive chemical, biological and pharmacological data. The final objective was to integrate the standardized small molecule datasets into their internal AI/ML platform for algorithm training, towards virtual hit-identification.

Our approach

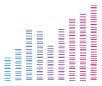
Excelra's Global Online Structure Activity Relationship Database (GOSTAR) provides a 360-degree view of million compounds, linking their chemical structure to biological, pharmacological and therapeutic information.

The heterogeneous and unstructured data captured from various data sources is transformed into a structured relational database format in GOSTAR. All the content in GOSTAR is captured manually and passes through a 3-step quality control process.

These normalized and structured datasets covering structure activity relationship (SAR), physicochemical properties, and ADMET parameters were integrated into the client's internal platform to train the AI/ML algorithms for model building and activity/property prediction to support hit identification and lead optimization.



Our contribution



Biological datasets

Biological data provides insights to understand the underlying mechanisms associated with disease state, prediction and validation of potential target proteins for the treatment and development of new bioassay techniques. Biological datasets within GOSTAR include protein/targets names, target family information, target synonyms, mechanism of action, target mutation information (deletions and substitutions) & binding affinity information.

Chemistry datasets

Chemistry datasets are useful in the design of high-throughput screening libraries which assist in identifying and validating therapeutic targets in silico. Chemistry datasets within GOSTAR include chemical structural representations, chemical line notations or identifiers (SMILES & InChI), molecular property descriptors, topological descriptors, topographical descriptors, structure-activity-relationships (SAR) & compound specific biological data.

Pharmacological datasets

Pharmacological data in drug discovery provides information about the compounds or drugs tested in animal models in combination with assay data on protein targets in cell- or tissue- based models. Pharmacological datasets withing GOSTAR include adsorption, distribution, metabolism, elimination and toxicity (ADMET) data, functional in-vitro assay & in-vivo assay properties.

Therapeutic datasets

The therapeutic datasets in drug discovery provide the valuable information in relation to the patient data. Therapeutic datasets within GOSTAR include indication names, safety and efficacy data, clinical/drug status information, dose information and adverse events or side-effects information.

High quality annotated datasets

GOSTAR provides a clear separation and structure to the data fields that can be easily imported into a database or graph structure. GOSTAR data is tagged to standard identifiers (such as Entrez gene ids or UniProt protein identifiers or ICD 10 disease classification) and the use of controlled vocabularies enables much simpler data integration from heterogeneous sources.

Flexible data delivery

GOSTAR data can be delivered to clients in the following file formats: relational database format (Oracle, PostgreSQL, MySQL), flat file or spread sheet formats (CSV, TSV, XML, XLS), chemistry specific formats (SDF, RDF) & semantic web formats (RDF, Turtle).

Where data means more

excelra

BOSTON | UTRECHT | HYDERABAD

Connect with our experts: marketing@excelra.com

www.excelra.com