The Bioinformatics Powerhouse Playbook

Harnessing the Power of Data to Accelerate Drug Development

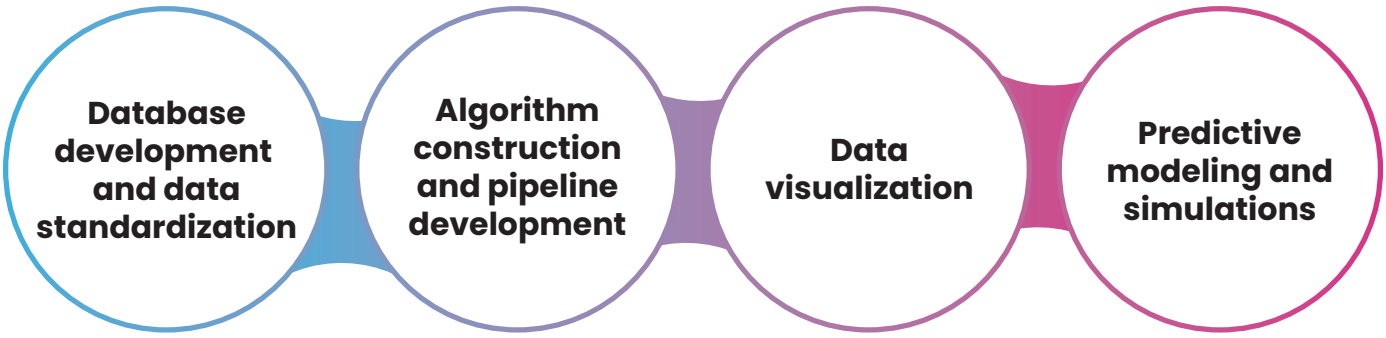
The Bioinformatics™ POWERHOUSE

excelra

The Data and Bioinformatics Revolution

Information is the backbone of innovation, and nowhere is this truer than in the field of drug discovery and development. As new technologies emerge and datasets grow in size and complexity, it becomes increasingly challenging for researchers and organizations to harness this information effectively. This is where the expertise of bioinformatics professionals can make all the difference.

This book serves as a definitive guidebook on the diverse applications of bioinformatics throughout the entire drug discovery and development process. From big data utilization to the integration of artificial intelligence and machine learning techniques, this comprehensive guide explores the numerous ways in which bioinformatics can revolutionize the field. By harnessing the full potential of bioinformatics, the pace of innovation in drug discovery and development can be accelerated, leading to better outcomes for patients. We dedicate this book to all of the leaders in the pharmaceutical and biotechnology industries who strive to remain informed and empowered in the face of an ever-evolving landscape.



Excelra: The Bioinformatics Powerhouse accelerating drug development

Excelra is an international powerhouse in consulting-led bioinformatics, with expertise in artificial intelligence, machine learning, and omics and next-generation sequencing data analysis. Our interdisciplinary teams of biocurators, bioinformaticians, data scientists, and engineers deliver harmonized data, optimized analysis pipelines, and advanced applications to accelerate drug discovery and development, making us a partner of choice to the world's top pharmaceutical and biotech companies.

Our biocurators and data scientists bring order to the chaos of data through expert curation, metadata management, ontology management, semantics integration, and knowledge graphs. Our bioinformaticians develop optimized genetic, multi-omics, and next-generation sequencing (NGS) data analysis pipelines. Our data engineers interpret biological information with predictive analytics, AI/ML models, and biostatistics. And our developers build cloud computing applications and visualizations to streamline collaborative research and decision-making.

This synergy of domain and technology help extract maximum insights for our clients in every phase of drug development as described in the following chapters.

What our Clients say about us

"We recently completed a project with Excelra on identifying and prioritizing target opportunities for SLE. The project was delivered successfully and on time. The Excelra team understood and applied our target filtering priorities and responded quickly and precisely to our questions throughout the project. They worked collaboratively and transparently and delivered meaningful data and insights to support our objectives and accelerate our research. We've been impressed by Excelra's professional and qualified team and we're looking forward to working with them again in the future."

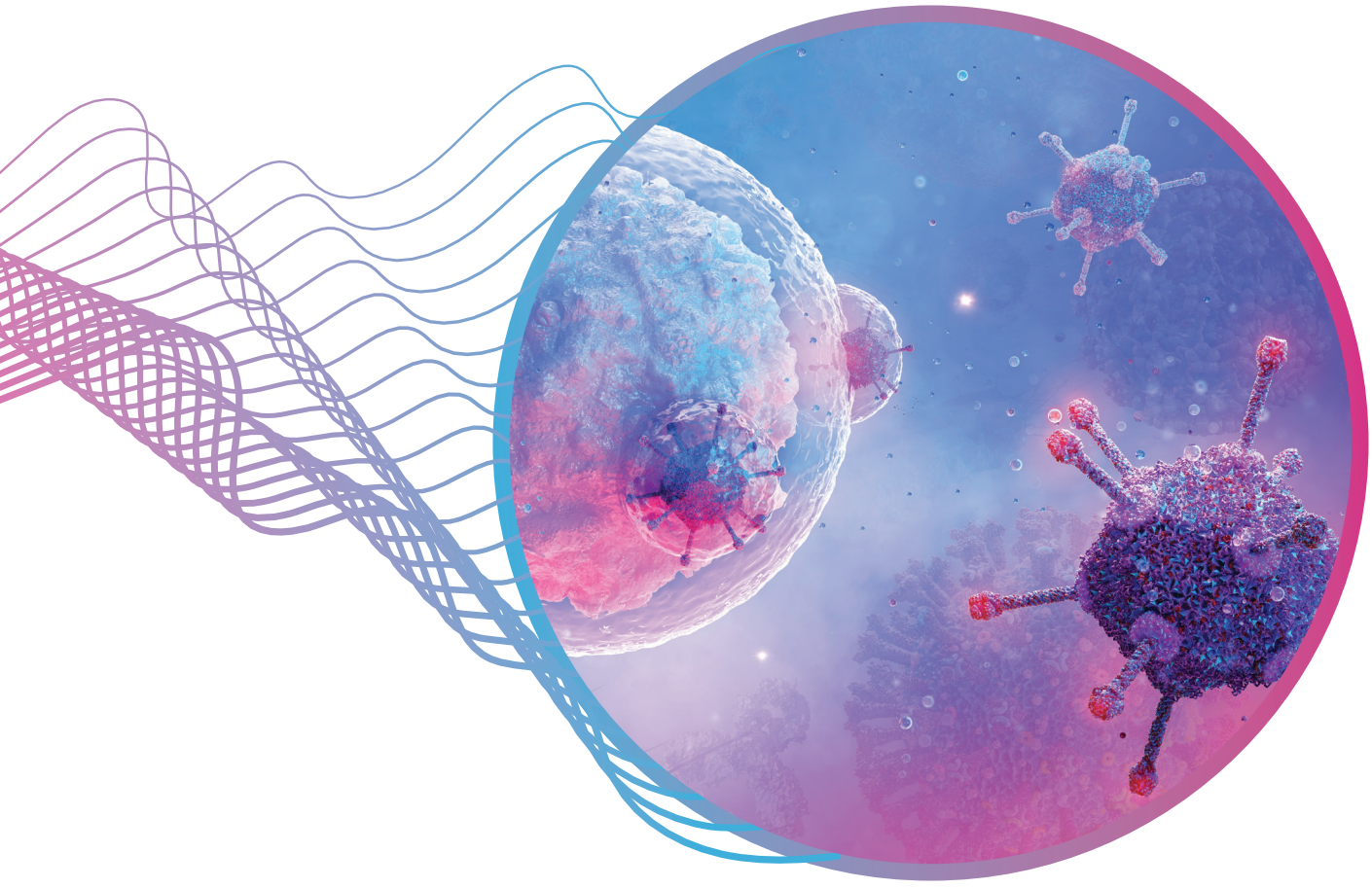
Dr Francois G. Gervais

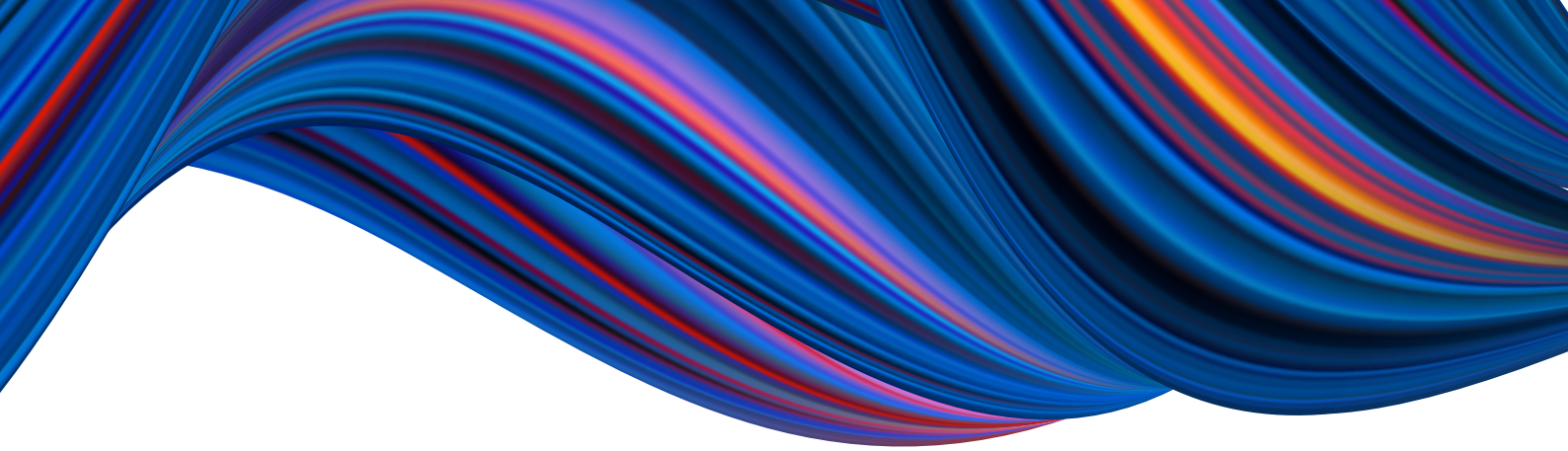
Executive Director, Discovery Biology

Ventus THERAPEUTICS

In a world, where the cost and complexity of developing novel medicines is ever-increasing, we help our clients make faster pipeline decisions, de-risk strategies, and accelerate the discovery of novel medicines for novel drug targets.

Index





Chapter 1

Disease landscape

The cornerstone for successful drug discovery and development

By integrating and analyzing large datasets, we have helped our client partners to understand disease landscapes and helped them position their drug development programs to be more competitive. We use computational analyses of omics datasets to identify, prioritize, and de-risk potential drug targets at the earliest stage of drug discovery and have helped uncover novel drug targets and pathways that can be targeted to develop novel medicines. Using computational tools, we have helped validate targets and their relevance to particular disease indications. We do this by assessing target gene expression, protein interactions, and pathway analyses, providing a comprehensive understanding of the target's role in disease mechanisms.

Case studies

1. Creating disease landscape for hyperphosphatemia
2. Identifying and prioritizing compounds for the treatment of rare monogenic blood disorders
3. Data-driven competitive landscape analysis to facilitate go/no-go decision in clinical development
4. Landscape survey of NSCLC for IP
5. Comprehensive analysis of putative drug targets and their comparison
6. Re-analysing database for novel targets

Disease landscape

Pharmaceutical and biotech companies start their study of disease by evaluating existing data. This involves reviewing past attempts to cure or combat the disease and analyzing market conditions to identify commercial opportunities. To do this efficiently, companies commission a disease landscape. A disease landscape is a compilation of all available information related to a specific disease and its treatments, including pre-clinical, clinical, and commercial data. It provides crucial insights for companies to decide whether to proceed with research and supports decision-making throughout the development program.

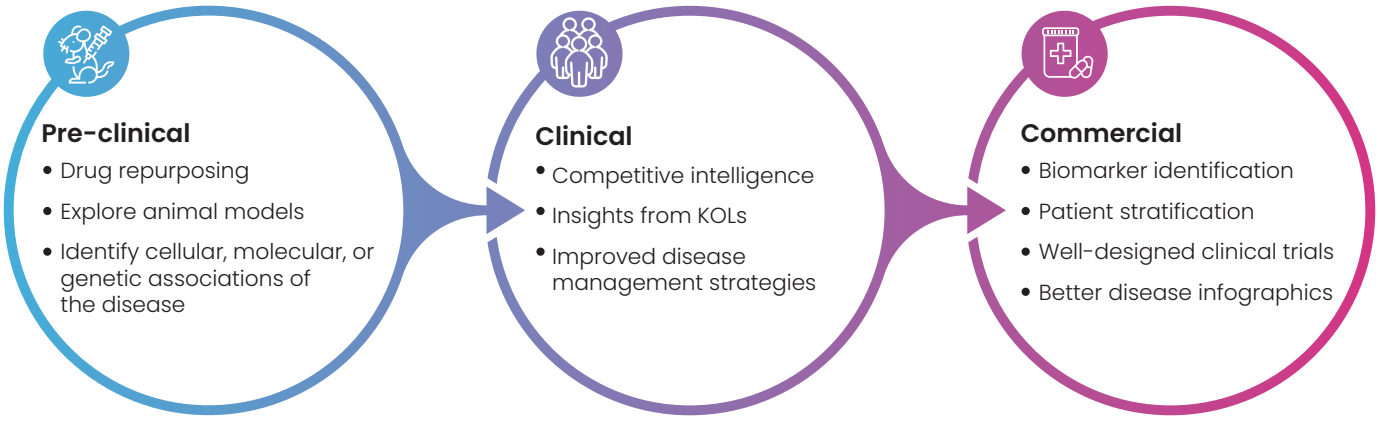


Figure 1: Key benefits of disease landscape assessments

When faced with critical go/no-go decisions, R&D teams evaluate disease landscape data from both scientific and strategic perspectives to gauge the potential success of a drug development program. The scientific assessment involves analyzing data related to the disease's biology, pathophysiology, treatments, biomarkers, genes, pre-clinical models, approved drugs, toxicity, safety, and mechanism of action (MoA). On the other hand, the strategic assessment focuses on market opportunities, commercial competitors, existing patents, and unmet patient needs. If either assessment indicates a low probability of success, the program is paused, recalibrated, or terminated.

Given the high stakes, disease landscapes must provide precise and high-quality data that fulfills all the evaluation requirements and help make critical go/no-go decisions.

Getting reliable information requires both domain expertise and data-handling expertise

Sourcing, extracting, normalizing, and analyzing data on disease biology in the scientific assessment is challenging for two major reasons:

- Data is non-uniform: The data produced by a broad coalition of scientific disciplines is naturally heterogeneous, inconsistent, and often incompatible.
- Data is scattered: It is often stored in different formats across disparate databases and requires distinct methods of analysis.

Research institutions and pharmaceutical companies face the daunting task of sourcing and selecting relevant data from a vast and open landscape and presenting it in an easily understandable format.

To address this challenge, they seek external partners to provide disease landscapes for drug development programs. Unfortunately, the results are not always optimal.

Often, even if data service providers successfully meet the standards required for data extraction and normalization, data quality fluctuates significantly and coverage data is inadequate in the absence of scientific domain expertise.

Conversely, suppliers with scientific credentials can select relevant, verified data but often lack the data-handling expertise needed to execute requests quickly while maintaining data consistency at high volume.

Companies who don't want to risk making wrong decisions due to inadequate or inconsistent data must partner up with external providers who assure data fluency and scientific expertise.

Besides analyzing the data on disease biology, companies compare their proposed compound against others already available. They do this by assessing the treatment regimens of all known approved drugs in circulation.

Treatment regime assessments incorporate analyses of a drug's mechanism of action, interaction with cellular targets, drug metabolism, efficacy, toxicity, and chemistry. Disease landscapes include all of this data, alongside information about previous and ongoing clinical trials, investigational compounds, animal models, unmet medical needs, and key opinion leaders (KOLs). These additional elements concentrate a company's focus on the potential for commercial success – a crucial consideration before committing to portfolio expansion.

With in-depth, validated data compiled on disease biology and treatment regimes, pharmaceutical companies can strengthen the foundations of their research decisions and accelerate their programs. Disease landscapes facilitate seamless progression from translational research into drug discovery and development and through each stage of pre-clinical and clinical trials.

Partner for reliable 360 disease landscape analysis

Excelra has been supplying disease landscapes to the world's biggest pharmaceutical companies for ten years. Its disease landscape curation team includes biologists, chemists, and data analysts collaborating with each other and their clients to provide disease landscapes that deliver precisely what the research program requires. The size and flexibility of Excelra's team mean that clients can engage them on short projects supporting immediate near-term objectives or on extensive, long-term projects to support every stage of the client's drug development program.

Excelra's disease landscape services fall into five broad categories as shown below.

Disease dossier

A disease dossier is created using text-mining algorithms that extract information from medical literature and data sets related to a particular disease. To ensure quality, the extracted data is manually validated by subject matter experts. The dossier provides insights into the clinical and commercial value associated with the disease or drug of interest. Depending on the objective, the dossier can include the benefits and limitations of different drug regimens, patient responses, clinical trials, possible targets, and disease similarities. All the information is delivered in a clear, easy-to-follow document to facilitate effective searching and easy referencing.

Target identification

Target identification is an early-stage discovery activity conducted by pharmaceutical companies in the pursuit of first-in-class or best-in-class drugs for their disease of interest. A scientifically ideal target must be demonstrably effective, impact relevant downstream pathways to improve the patient's condition and/or achieve commercial viability by addressing unmet medical needs. Target identification exercises help researchers select targets most relevant to the pathophysiology of a disease and are conducted using gene or tissue expression, knock-out/knock-in studies, downstream pathways analysis, and safety assessments.

Disease centric repurposing

Pharmaceutical companies don't only focus on the development of new drugs. Sometimes, repurposing existing drugs is a more effective and efficient approach to combating the causes or symptoms of a disease. Disease-centric repurposing requires

the identification of drugs that could be used on alternative indications. Candidates for repurposing are selected by analyzing data on their mechanism of action, efficacies, target assessment, and safety analysis.

Indication prioritization

Advanced bioinformatics techniques and chemical, biological, and clinical intelligence can be applied to high-quality, manually curated data to reveal the most relevant indications for a drug or target under study. These indications are then evaluated and prioritized in order of success probability. The prioritization of indications can accelerate the drug development process and reduce the number of failures.

Biomarker identification

Biomarkers are relevant to the entire research program. They help map a disease's progression, support patient stratification, contribute to the identification of perturbed pathways, and facilitate the correlation of mechanisms of action. Biomarker identification is, therefore, an essential stage of translational medicine. Using extensive omics data sets and advanced data-mining techniques, Excelra's experts provide detailed insights on biomarkers associated with a given disease.

Exclera is the standout partner to deliver disease landscapes, given its unique combination of scientific domain expertise and data analysis capabilities. Our disease landscapes are invaluable resources to support short-term objectives or long-term projects. See our case studies for specific examples.

Case study 1
Creating disease landscape for hyperphosphatemia

Client's challenge and goal

Clinical trial data helps pharmaceutical companies make informed decisions along their drug discovery journey. A biotech firm engaged Excelra to collect information on clinical trials associated with hyperphosphatemia to assist in its ongoing research program into that disease. The client requested a comprehensive disease landscape, including all available information on drug doses, efficacy, and endpoints.

Figure 2: Concepts explored for hyperphosphatemia disease landscape

Our approach

Excelra's team collected the required information via in-depth data mining on CT.gov, JMACTR, UMIN-CTR, and other relevant repositories. Those clinical trials that met the client's criteria were shortlisted for curation by Excelra's subject matter experts, who extracted efficiency, dose, and endpoints data. The data were analyzed and prepared for downstream processing, and delivered to the client to the standard required.

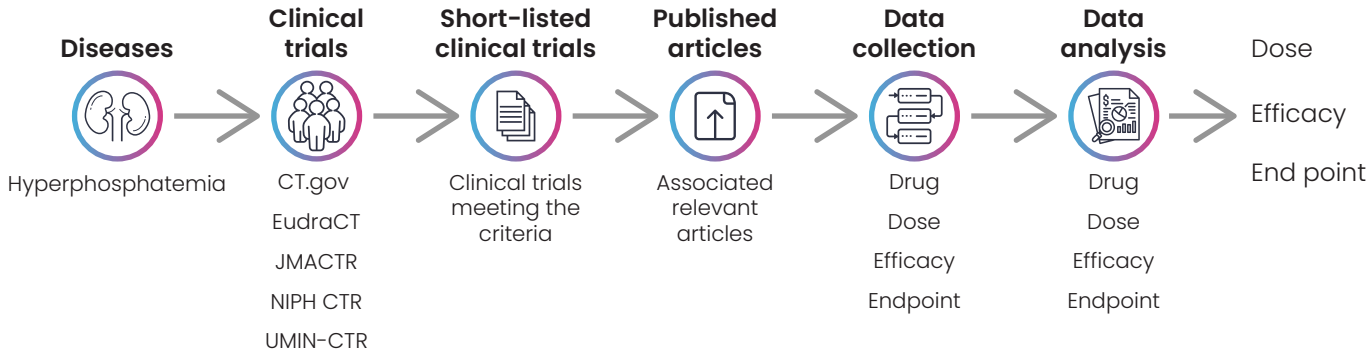


Figure 3: Efficacy, dose, and endpoint profiling of drugs associated with hyperphosphatemia

Excelra also prepared a comprehensive landscape of phosphate binders. The pill burden and adverse effects (particularly gastrointestinal intolerance) associated with phosphate binders often contribute to poor medication adherence.

At the end of Excelra's investigation into phosphate binders, data was revealed relating to indications, reported efficacy, dosages, and physicochemical properties.

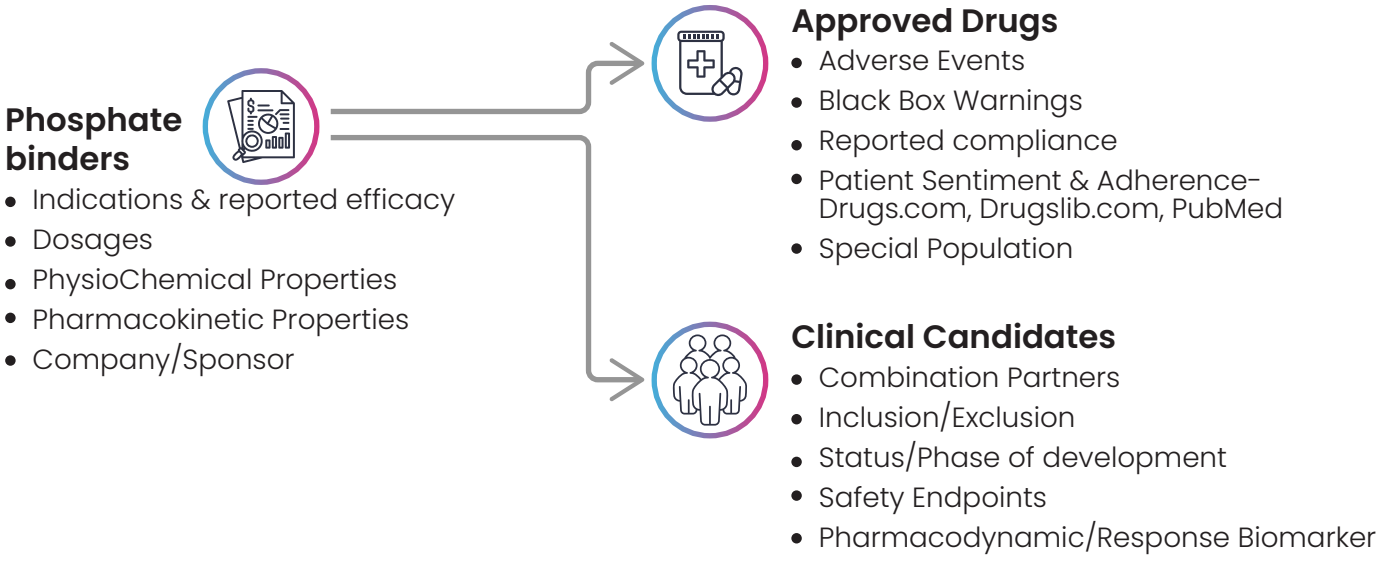


Figure 4: Comprehensive landscaping of phosphate binders

Result and impact

The client received comprehensive data on clinical studies associated with hyperphosphatemia and phosphate binders, delivering valuable insight into dose efficacy and endpoint relationships.

Case study 2

Identifying and prioritizing compounds for the treatment of rare monogenic blood disorders

Client's challenge and goal

A small pharma company requested Excelra's help to identify potentially effective drugs for the treatment of rare monogenic blood disorders. Rare monogenic disorders are primarily caused by single gene mutations, so Excelra explored existing literature and datasets to establish disease-drug correlations. Many approaches were used, including disease similarities, drug-gene signatures, and genome-wide association studies (GWAS). The extracted information was used to create disease-drug pairs based on their mechanism of action

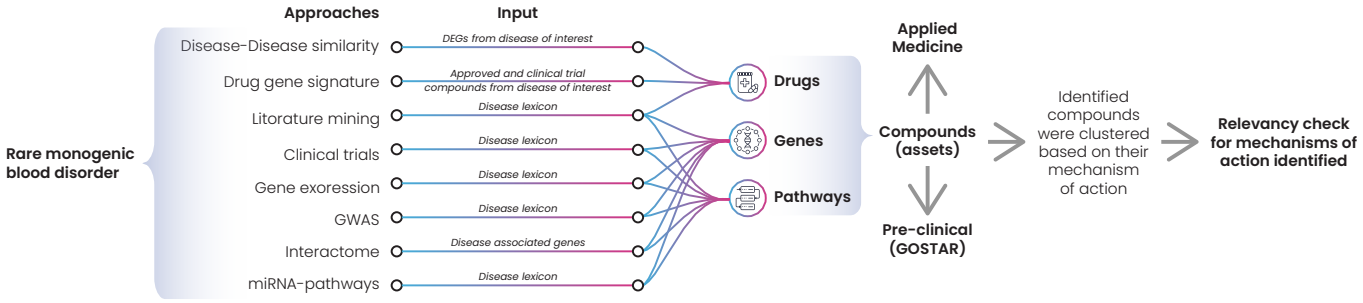


Figure 5: Approaches for asset identification in relation to rare monogenic blood disorders

Our approach

The top five mechanisms of action were selected, and drugs were prioritized for each of them. An in-depth analysis of each of the MoAs was completed considering the following points:

- Relevance of targets and MoAs in the disease
- Clinical or pre-clinical scientific evidence
- Known literature on animal models, target safety, and hypotheses availability

Following this stage, the best drugs for each mechanism of action were recommended to the client, and the MoAs and compounds were prioritized according to Excelra's prioritization process (Fig.10). Excelra conducted a further relevancy check on the mechanisms of action for compounds/assets to discover if they promoted or alleviated disease and complications

.

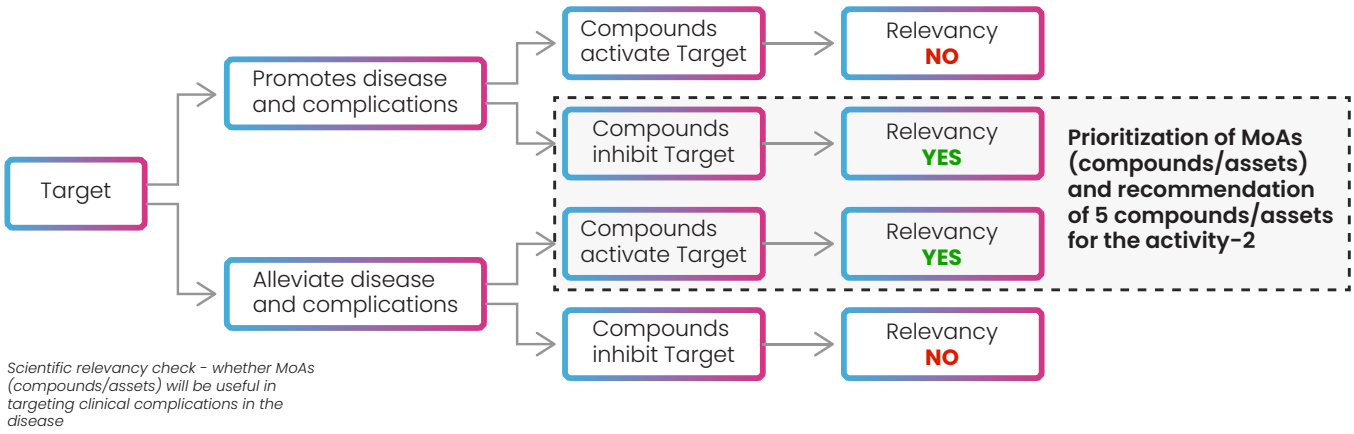


Figure 6: Prioritization of MoAs and compounds/assets

Result and impact

Our client received:

- a scientific rationale for each MoA in the context of disease pathophysiology
- a priority list of drugs and recommendations for each disease
- recommendations regarding suitable animal models for PoC experiments

Case study 3

Data-driven competitive landscape analysis to facilitate go/no-go decision in clinical development

Client's challenge and goal

A Switzerland-based large pharma company engaged in the development of novel antibody therapeutics against rheumatoid arthritis (RA), was analyzing the data to demonstrate the advantage of longitudinal meta-analysis over conventional meta-analysis that uses end-of-study (EOS) data, toward facilitating more effective model-informed drug development (MIDD) decisions.

The objective of the analysis was to determine the competitive position of the client's novel antibody in early clinical Phase II B versus all the approved biologics against rheumatoid arthritis (RA). The client was mainly interested in performing a quantitative assessment of the longitudinal time course of clinical efficacy that would enable informed decision-making on further clinical development.

The client approached Excelra to develop a model-based meta-analysis (MBMA)-ready dataset, by curating all the existing scientific evidence around the efficacy of marketed biologics for RA.

We received the following requirements:

- the curated dataset includes a summary time-course responses on clinical outcomes used in late-stage clinical trials
- the data covers information about prior and concomitant medications including:
  - respective category-wise percentage of patients (with response status to medications)
  - baseline patient characteristics and sample size including N in statistical analysis

Our approach

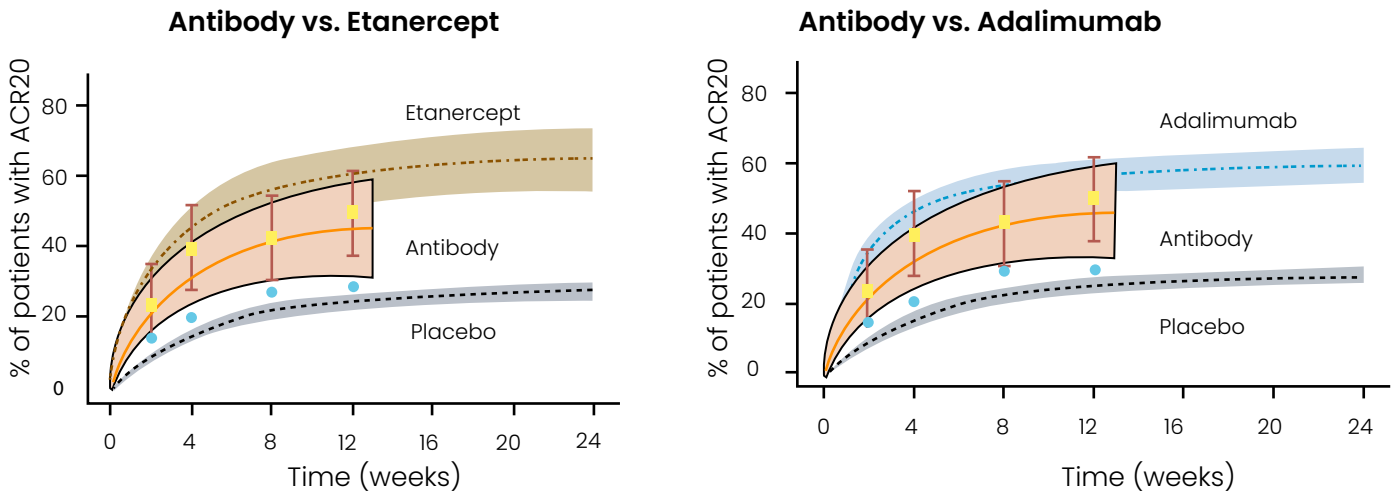
In line with the specified requirements, Excelra's Clinical Pharmacology group used a robust scientific curation methodology coupled with systematic literature review (SLR), to synthesize data on existing therapeutics for performing model-based meta-analysis (MBMA).

We completed the following steps::

- Defined project scope with PICOS methodology for conducting systematic literature review in PubMed
- Screened, labeled, and developed a database for enabling further qualification and selection of relevant publications according to PICOS specifications
- Identified additional references following a thorough search across FDA drug labeling information and traditional meta-analysis publications (119 sources identified)
- Developed a customized clinical outcomes database, capturing:
  - Clinical outcomes summary data (Time vs response)
  - Patient population details (Baseline characteristics, prior and concomitant therapy)
  - Interventions (Dose regimen)
  - Comparator (Dose regimen)
  - Study design (Sample size)
  - A rigorous 3-level quality control (QC) process was employed for database development

Result and impact

Based on the custom datasets developed by Excelra, the client was able to demonstrate the advantage of longitudinal data analysis over conventional EOS meta-analysis. Combining the resultant longitudinal MBMA on late-stage clinical outcome ACR20, with inhouse Phase II B data of the novel antibody, helped the client to make a well-informed, data-driven 'no-go' decision for further clinical development of the biologic against RA.



. Figure 7: Example of an actionable finding from Excelra's analysis – because of its inferior efficacy profile in RA (ACR20), the novel antibody has lower chances of success vs competitor drugs, Etanercept and Adalimumab.

After refining by our team, the client's in-house literature database (clinical trial outcomes database) for rheumatoid arthritis included 37 Phase II and III studies describing 13474 patients, 75 arms, and 502 summary points.

We updated the database with each time point, which was data-digitized from the illustrative time course curve in each study. This enabled the client to compare the antibody of interest with the available biologics for Rheumatoid Arthritis.

The magnitude of response and the associated time course analysis from Excelra's databases showed that the novel antibody had lower chances of success owing to its inferior efficacy profile in RA (ACR20) when compared to competitor drugs, Etanercept and Adalimumab (as shown in the figure below)

Case study 4

Data-driven competitive landscape analysis to facilitate go/no-go decision in clinical development

Client's challenge and goal

Our client is a large pharmaceutical company based in the United States. One of its research teams is seeking potential new treatments for non-small-cell lung cancer and is engaged in identifying appropriate biomarkers to help design effective clinical trials.

The identification process requires a thorough survey and analysis of existing NSCLC clinical trial data. The data needs to be extracted from a wide library of literature and consistently structured before analysis. This process demands significant time and resources, so the client engaged our team to execute the survey, extract the data, and deliver a comprehensive analysis.

The goal of the project was a detailed report on clinical studies with checkpoint inhibitors for NSCLC that target PD-L1 or PD-1. The report also needed to include a detailed review of the inhibitors' potential benefits within four different lines of therapy: naïve patients, first-line therapy, second-line therapy, and third-line therapy.

Our approach

To meet the client's requirements, we developed a text-mining algorithm to identify relevant literature. Once the algorithm had delivered an exhaustive list, our scientists manually curated, annotated, and analyzed the information, delivering a refined list to the client.

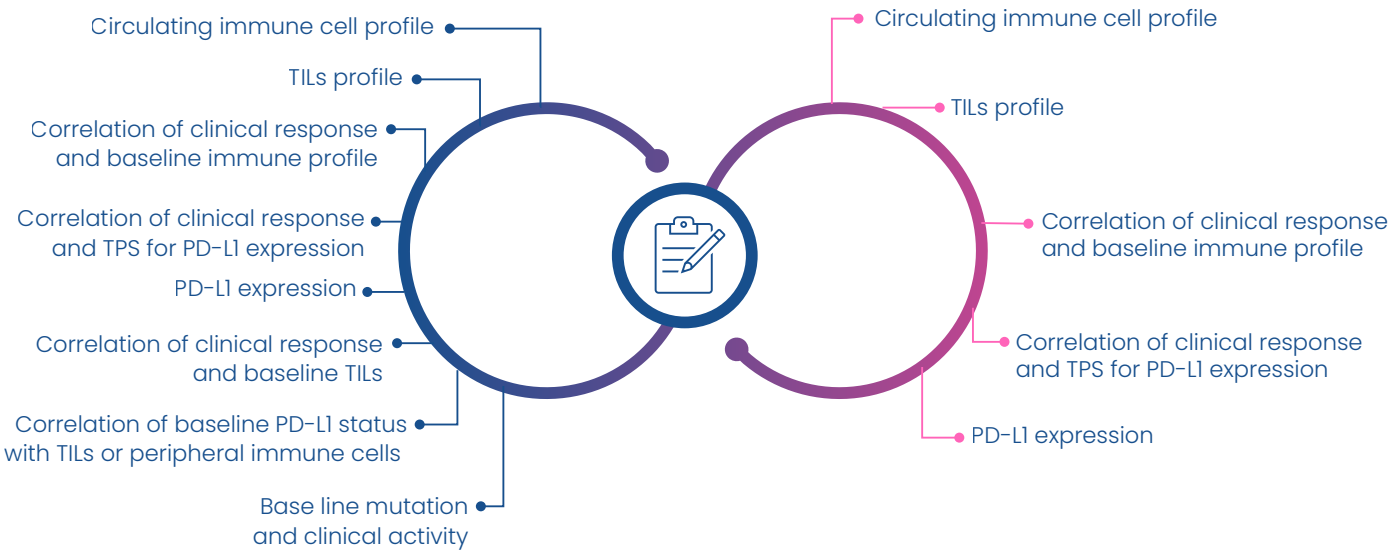


Figure 8: Anti-PD-1 and anti-PD-L1 treatment and efficacy data points.

The curation process captured information from four focus areas:

1. Immunotherapies NSCLC clinical studies. Our qualified subject matter experts curated 82 relevant studies. The compiled NSCLC landscape included details of drugs, lines of therapies, and efficacy-related data points.

2. Comparative analysis of anti-PD-1 and anti-PD-L1 treatment and efficacy endpoints. We identified data points common for both anti-PD-1 and anti-PD-L1, as well as data points unique to anti-PD-1

3. Treatment regimens in NSCLC across different lines of therapy and comparative analysis of efficacy endpoints.

4. PD-L1 Assays and association with the clinical response across different lines of therapy.

Results and impact

Our team delivered high-quality information quickly and according to precise specifications.

With our comprehensive data, the client was able to:

- select relevant biomarkers
- develop a greater understanding of the mechanism of immune response
- identify alterations in immune cell/TIL profiles in different lines of therapy

Overall, our analysis-ready data has supported the client's clinical trial decisions and is contributing to its ongoing research.

Case study 5

Comprehensive analysis of putative drug targets and their comparison

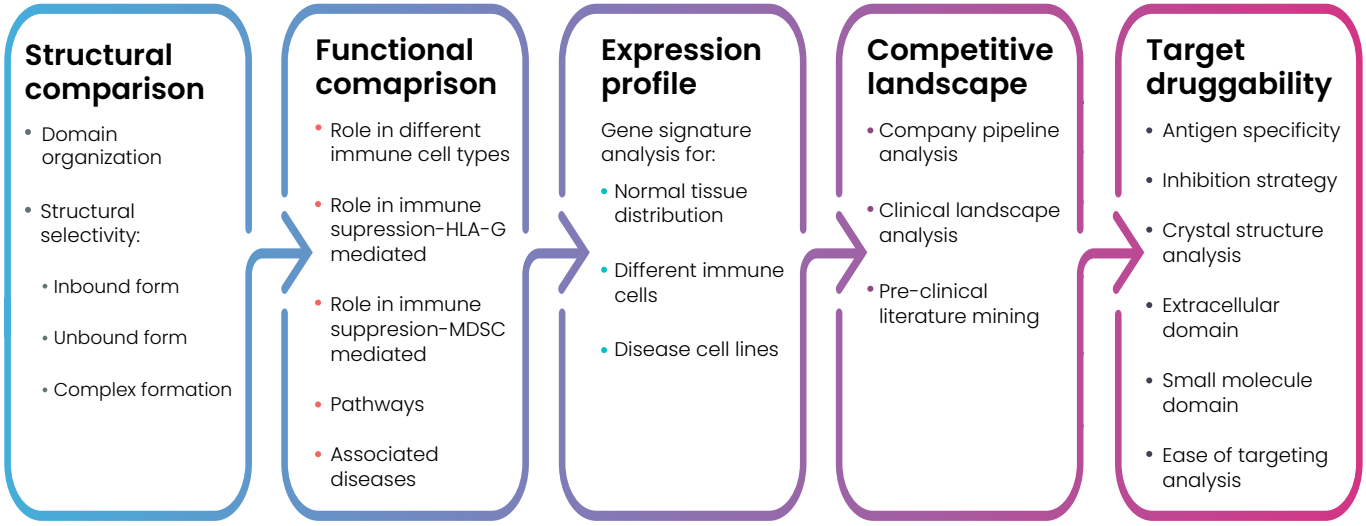
Client's challenge and goal

A large pharma company based in the United States requested a comprehensive assessment and comparison of two new drug targets of interest which are the most abundantly expressed inhibitory leukocyte immunoglobulin (Ig)-like receptors.

The client's goal was to facilitate its early discovery phase decisions using the insights gained from the analyses.

Our approach

Our team performed the following analyses:



Results and impact

The client was able to facilitate its decisions in the early discovery phase thanks to the insights we uncovered for the following aspects of the targets:

- Target function in healthy tissues as well as its association with disease(s)
- Molecular pharmacology, target expression across human tissues/organs, species, gene alterations, and target interactions with other proteins/genes
- Sequence, crystal structure, domains, and homology modeling
- Adverse events or toxicity data for compounds at any stage of development
- Recommendation of animal models and pre-clinical assays
- ON- and OFF-target safety assessment and de-risking strategies

- Competitor profile: Competitive landscape of drug development by stage (approved, clinical, and pre-clinical drugs)

We also recommended the preferred target to proceed with. One reason was that one of the targets would be a first-in-class therapy with no current evidence of clinical testing in humans. Another argument was related to the ease of targeting based on crystal structure analysis.

Case study 6

Comprehensive analysis of putative drug targets and their comparison

Client's challenge and goal

A customer was interested in finding new potential targets for siRNAs based on their existing databases.

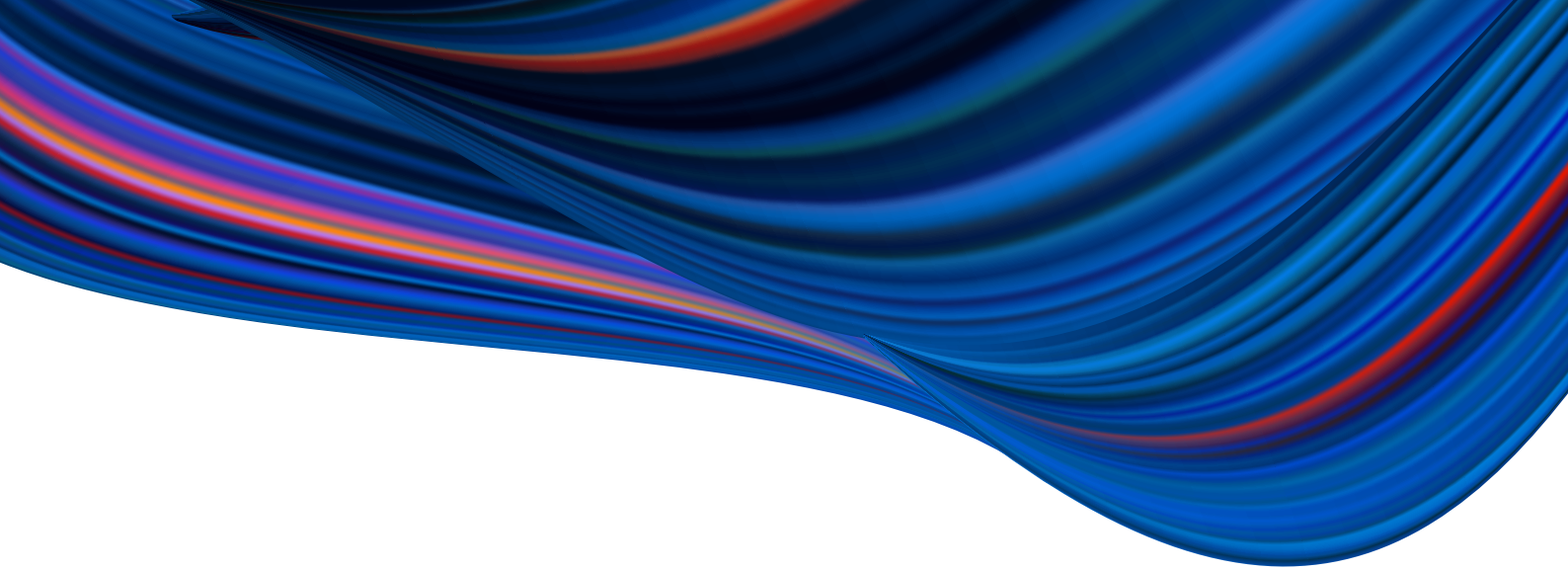
Our approach

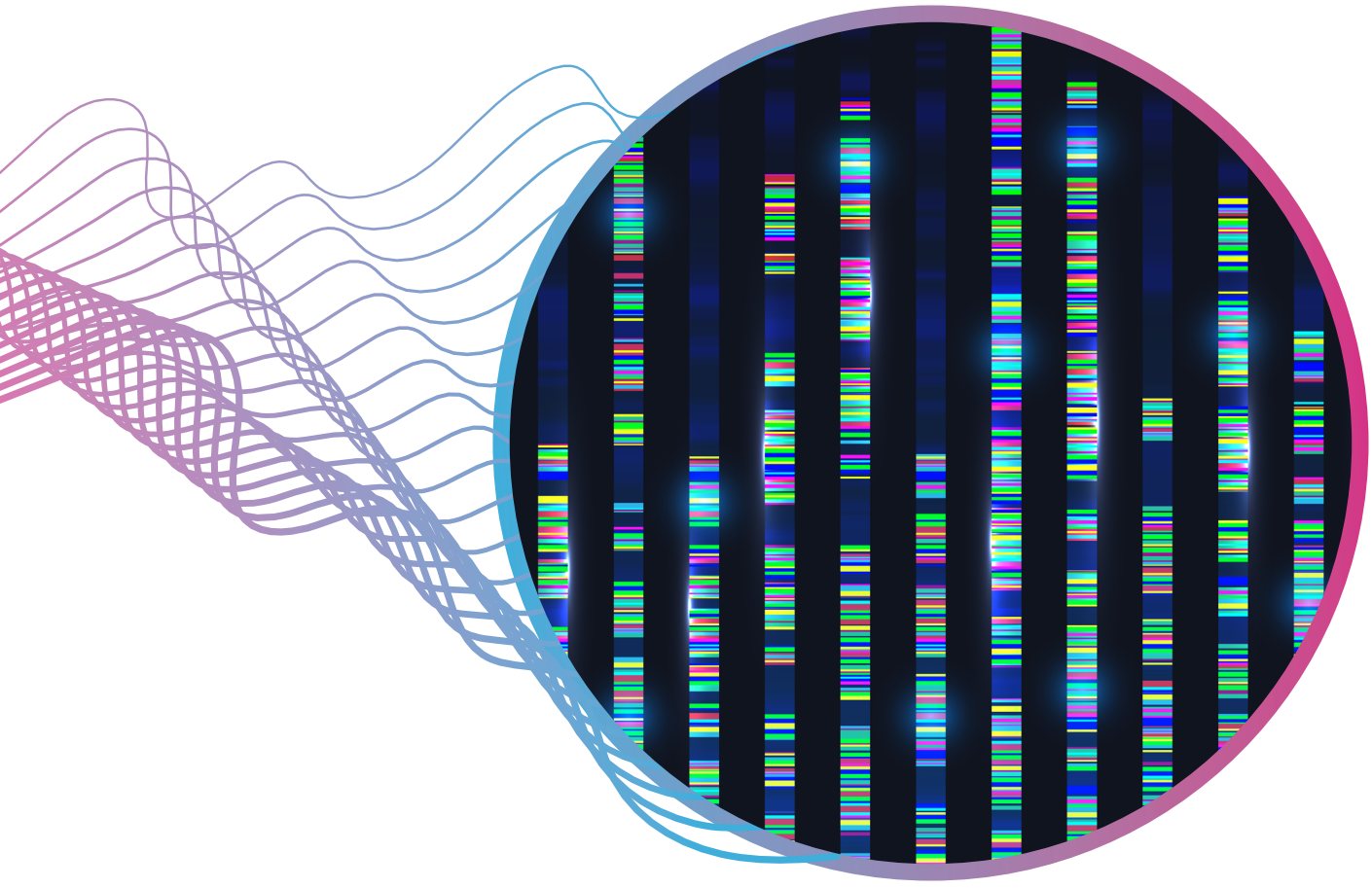
Our team performed a study combining the following analyses:

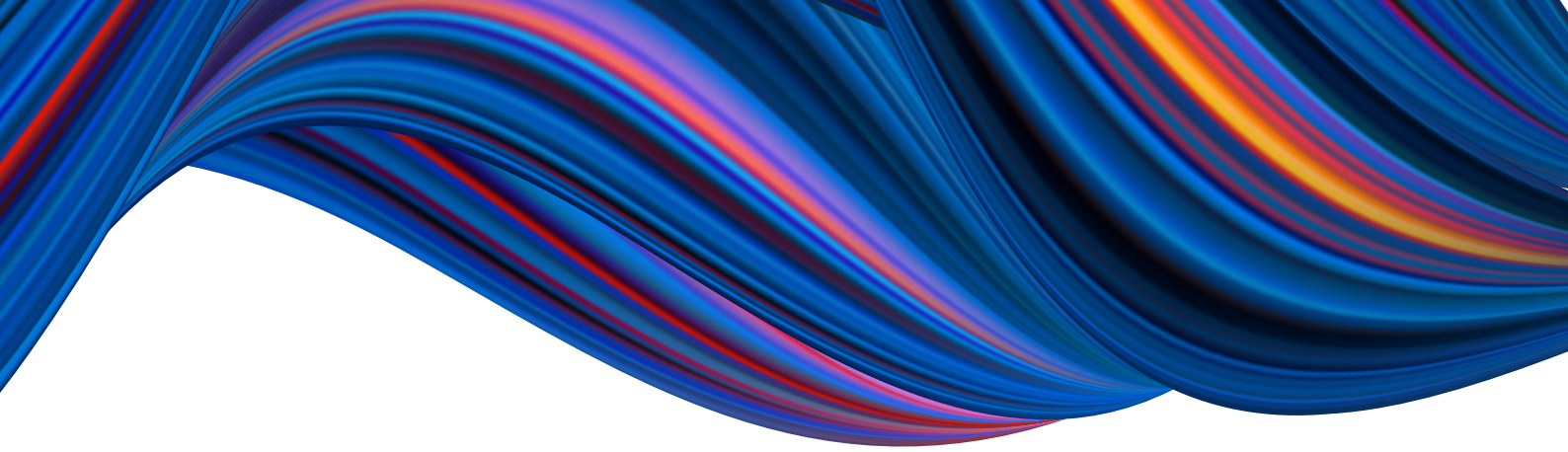
- siRNA hit calling and determining on/off-targets
- Selection of siRNAs for subsequent toxicity screen
- Define thresholds for cell viability, toxicity, household gene stability

Results and impact

We identified compelling on/off-target interactions with limited effects on cell viability and toxicity.







Chapter 2

Preclinical hit ID, lead discovery, and optimization

Techniques such as virtual screening, molecular docking, and quantitative structure-activity relationship (QSAR) modeling help identify and optimize drug candidates. These methods enable researchers to evaluate large libraries of compounds and accurately predict their binding affinity to a target. With these insights, they can refine the list of potential drug candidates for further experimental validation.

Case studies

1. Comprehensive analysis of putative drug targets and their comparison
2. KRAS program IP development using data curation

Case study 1

Comprehensive analysis of putative drug targets and their comparison

Client's challenge and goal

Our client is a European pharma company focused on discovering and developing small-molecule medicines with novel modes of action. A key stage of their research demands the processing of sequenced data through a next-generation sequencing (NGS) pipeline.

The client's DNA-encoded library (DEL) had a large volume of data but low compound coverage (about 3-10 counts per compound). A pragmatic statistical approach using differential gene expression was required to reliably detect true positives and avoid the obstacles caused by the data.

The goal of this project was to develop a hit-calling algorithm to find candidates for testing.

Our approach

With our data scientists supported by domain experts, we were able to conduct all the necessary preparatory research before building pipelines.

To ensure the algorithm was optimized for the specific data involved, we started with a thorough analysis of data from 72 samples with raw counts for each of the 5 million compounds in the client's library. Understanding that low raw count numbers increased the risk of selecting false positives or dropping false negatives, we put particular emphasis on normalizing the data set so that trends were maintained across all the generated data.

We reviewed various DEG algorithms based on the analyzed data and contrast sheet. After meticulous comparative analysis, it was clear that DESeq2 provided the most reliable and consistent results.

With the algorithm selected, we built an effective and efficient automated pipeline, containerized with NextFlow.

Finally, our data scientists rigorously tested the pipeline to ensure the outputs were generated in the form and according to the standard demanded by the client.

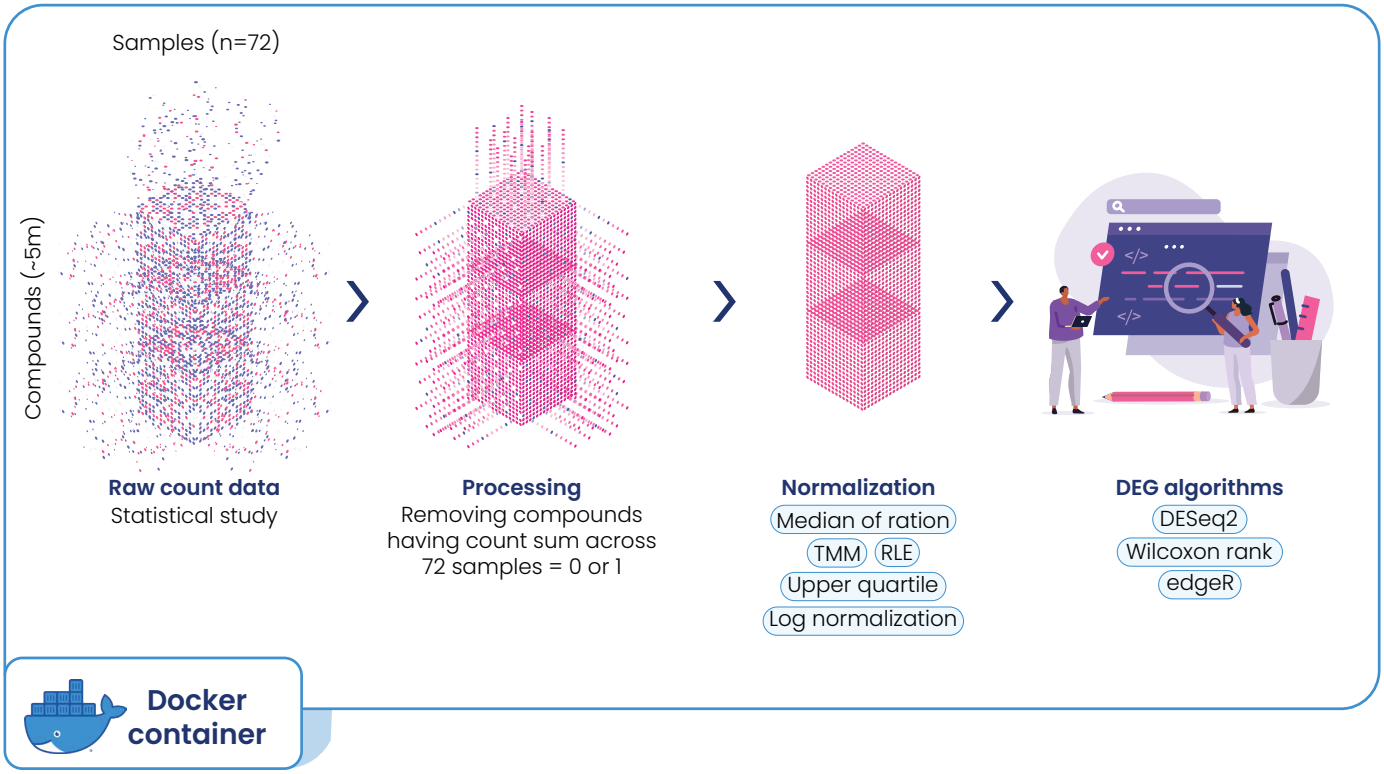


Figure: 1 Excelra's workflow from statistical data analysis to creating an efficient automated pipeline.

Results and impact

The client is able to accurately identify true hits from DEL selection output data even if the raw counts are low in numbers. This is possible thanks to the fully functional, deployable, portable NextFlow pipeline that our team delivered.

The client received moreover detailed documentation to facilitate a smooth transition from the Excelra's to client's team, and to ensure optimized performance and minimal downtime.

Case study 2

KRAS program IP development using data curation

Client's challenge and goal

Our client is a large US biopharmaceutical company engaged in Kirsten rat sarcoma viral oncogene homolog (KRAS) research. A foundational phase of their R&D activity involves collecting information on existing patents related to KRAS and data on all associated structures and activities.

With thousands of KRAS-related patents published every year, the process of gathering, reviewing, and extracting the relevant data demands, however, a significant investment of time and resources.

Knowing our capabilities, the client engaged our team to extract and validate the required content and deliver clean, consistent, and analysis-ready data.

Our approach

With over 60 PhDs in our data curation team, we have the domain expertise our client demanded to identify the relevant literature, extract the appropriate data, and deliver it in a standardized, analysis-ready format.

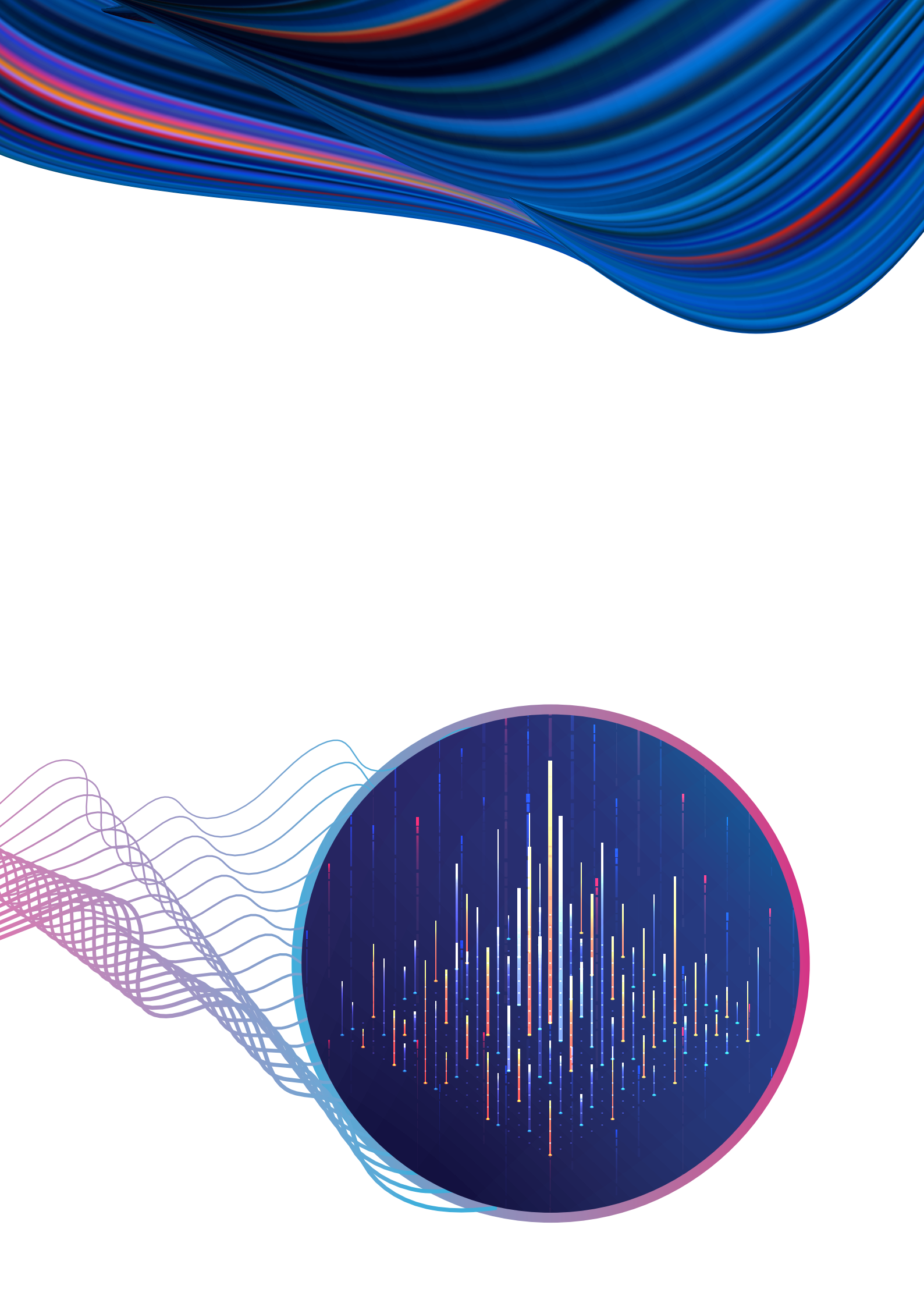
As a source of the ADME data, we chose Excera's GOSTART database. In comparison with another well-known database, ChEMBL, GOSTAR has not only more data but also the data diversity is superior to that of ChEMBL. If you are interested to learn the importance of this aspect in a drug development process go to ChEMBL vs. GOSTAR – Data diversity and compound coverage

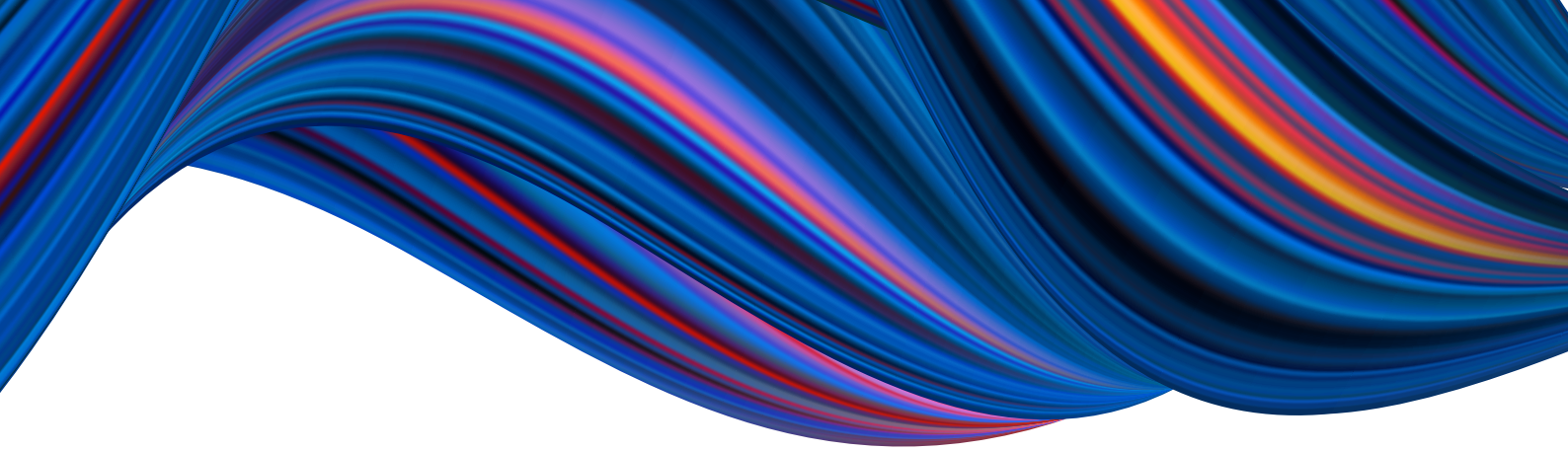
Our curation process includes three stages of data extraction: manual curation, review, and quality control. Thanks to a combination of our scientific expertise and technical excellence, we were able to collate, prepare and deliver the client's data set in an exceptionally short time. We ensured that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

Results and impact

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 continuously returned with similar data curation requests.





Chapter 3

Systems biology approach to assess mechanism of action (MoA), efficacy, and safety in drug development

We have helped our clients integrate various types of biological data to simulate the complex interactions between drugs, targets, and biological systems. These models can provide valuable insights into a drug's mechanism of action (MoA) and help predict drug responses, side effects, and drug-drug interactions. Furthermore, we are able to analyze and interpret preclinical genomic, transcriptomic, and proteomic data. This information is crucial for understanding drug efficacy, toxicity, and pharmacokinetics, ultimately aiding in pre-clinical development.

Case studies

1. Screening adverse-events-related data
2. Optimizing dose regimen for Paclitaxel

Case study 1
Screening adverse-events-related data

Client's challenge and goal

Our client is a European publishing house and owns a proprietary tool for preclinical toxicity, clinical, and pharmacovigilance studies. The goal of the project was to add to the tool accurate, up-to-date and validated content on drugs and targets to its adverse events. The update of the database needed to be facilitated by Excelra's data-mining, classification, and excerption services. The validated content would then be used to improve the sensitivity of the client's text-mining pipeline.

Our approach

Excelra's expert biocurators ran an in-depth, manual validation on the extracted PMIDs for adverse events. The validated PMIDs were then used to train the client's text-mining pipeline and fine-tune its output. To obtain the desired results, our team exhaustively screened approximately 1000 articles a day. The literature included: case reports of drug-induced adverse events; potential association between a drug (or drug class) and an adverse event; review of safety data, drug, and drug class; preclinical toxicology results with a new drug candidate or a known drug; knockdown/knockout studies; and articles correlating disease with a genomic finding.

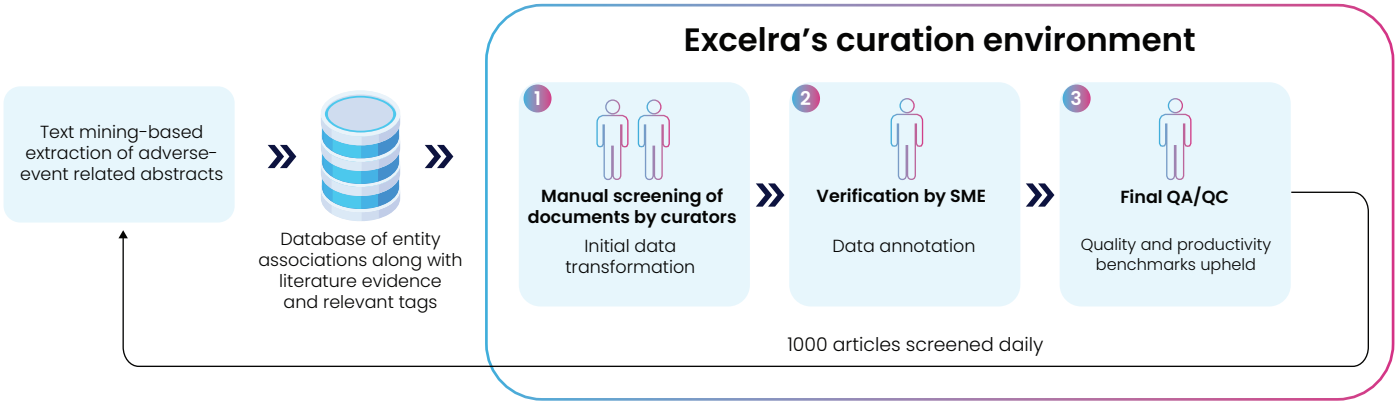


Figure 1: Excelra's data-curation pipeline

Result and impact

The validation provided to the client as a result of our verification and annotation exercise helped ensure the correct data was entered into their production environment. Our curation services helped the client successfully update the database twice a week with validated content. The additional entries on drugs and targets also helped to train the client's existing text-mining pipeline, substantially improving its efficiency

Case study 2
Optimizing dose regimen for Paclitaxel

Client's challenge and goal

A US-based biotech company was interested in quantifying a relationship between different Paclitaxel doses and regimens on safety and efficacy using summary-level data across published clinical trial literature by model-based meta-analysis (MBMA).

The goal of this project was to identify and build a database with safety and efficacy data for Paclitaxel monotherapy dosing regimens in clinical practice.

Our approach

- Gathered the scope with PICOS methodology for systematic literature search in PubMed
- Retrieved literature was scientifically screened and labeled with appropriate variables, following which a database was developed to ease further screening and selection of the most appropriate publications as per PICOS specifications
- After full-text examination, 55 publications describing 49 double-blind phases I, II and III clinical trials with paclitaxel monotherapy in multiple oncology indications were curated
- A clinical outcome database of these publications was developed by a sheer scientific approach to capture clinical outcomes summary data (time vs. response in terms of safety and efficacy endpoints) with all other scoped information for each data point, about patient population (indication), interventions (dose regimen), comparator (dosage regimen), outcomes, study design (sample size) details
- Process quality was maintained by peer review at each step of database development

Result and impact

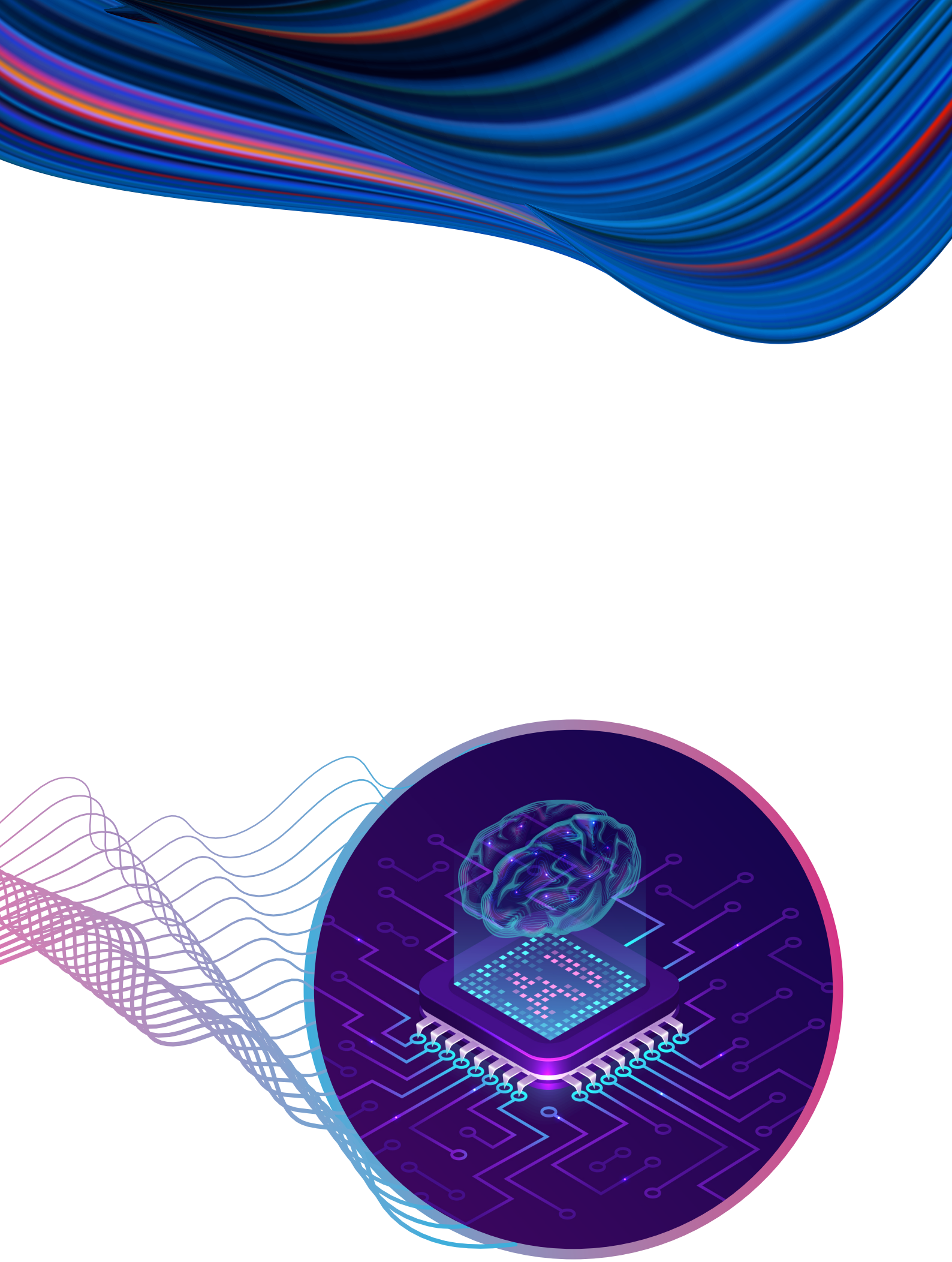
The client's MBMA on Excelra's data supported the choice of weekly (QW) over every 3-week regimen (Q3W) for the doses included in the modeling, for a better-balanced safety and efficacy profile

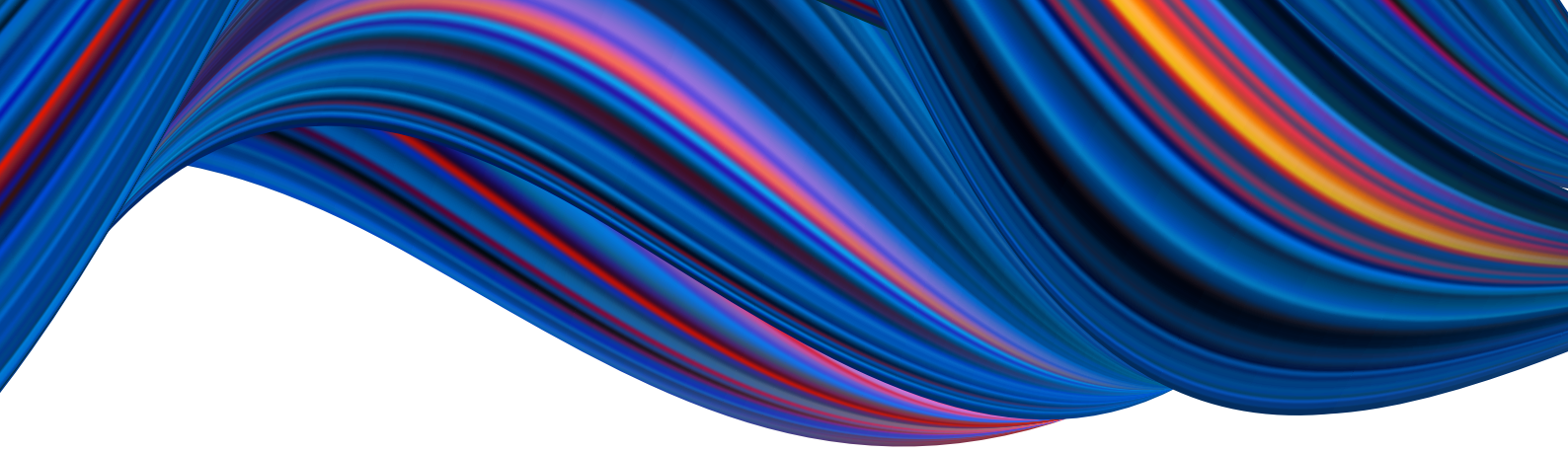
Our client received a refined, structured database i.e. drug-based database for multiple oncology indications (breast cancer, ovarian cancer, glioblastoma, lung cancer, and other mixed tumor types).

From each study, using scientific rationale we extracted the data on prespecified

essential outcomes that represent safety, efficacy, and their associated treatment and dose regimens.

Wherever reported, we captured efficacy endpoints from tables and digitized time-to-event curves of OS, PFS, etc. for each time point, or censored and recorded them in the database.





Chapter 4

AI/ML-based approaches to build predictive models in drug development

We help our clients engage in AI/ML methods to accelerate their drug discovery pipeline. Since the success of AI/ML models relies on the underlying datasets, we have helped clients by creating harmonized and standardized datasets that are fed into AI/ML models. We use our datasets to perform predictive modeling of ADME data or use pipelines with AI/ML technologies to predict 3D structures of proteins from cryo-EM experiments.

1. ChEMBL vs. GOSTAR – Data diversity and compound coverage
2. Selecting and preparing data for AI/ML predictive modeling

Case studies

1. Structured and analysis-ready data for AI/ML-based drug discovery
2. Combination feasibility prediction for checkpoint inhibitors for a biologic
3. Powering up cryo-EM data for faster drug discovery

Case study 1

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

Client's challenge and goal

A US-based 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.

The client required high-quality, harmonized, and structured datasets of small molecules, encompassing comprehensive chemical, biological, and pharmacological data.

The goal was to integrate the standardized small molecule datasets into their internal AI/ML platform for algorithm training, towards virtual hit-identification.

Our approach

We applied our Global Online Structure Activity Relationship Database (GOSTAR), which provides a 360-degree view of millions of compounds, linking their chemical structure to biological, pharmacological, and therapeutic information.

The heterogeneous and unstructured data captured from various data sources are 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 adsorption, distribution, metabolism, elimination, and toxicity (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.

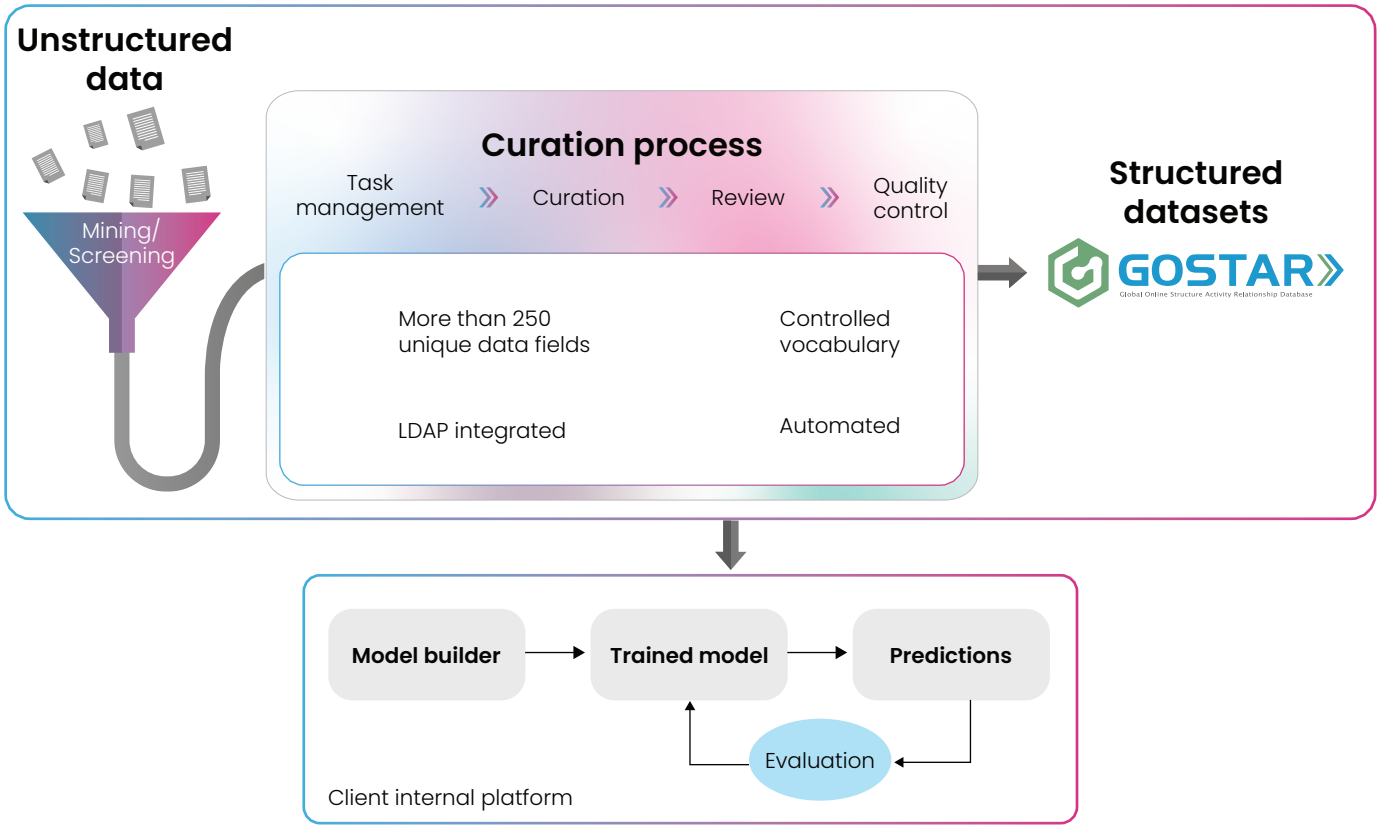


Figure 1: Structured datasets from GOSTAR are integrated into the client's internal platform to train AI/ML algorithms

Result and impact

The client received the following standardized datasets for small molecules:

Biological datasets

They help understand the underlying mechanisms associated with disease state, prediction, and validation of potential target proteins for the treatment and development of new bioassay techniques. Withing GOSTAR, these datasets include protein/target names, target family information, target synonyms, mechanism of action, target mutation information (deletions and substitutions) and binding affinity information.

Chemistry datasets

They 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 and InChI), molecular property descriptors, topological descriptors, topographical descriptors, structure-activity-relationships (SAR) and compound-specific biological data.

Pharmacological datasets

In drug discovery, these datasets provide 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 within GOSTAR include adsorption, distribution, metabolism, elimination, and toxicity (ADMET) data, functional in vitro assay and in vivo assay properties.

Therapeutic datasets

In drug discovery, these datasets provide valuable information in relation to 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 spreadsheet formats (CSV, TSV, XML, XLS), chemistry-specific formats (SDF, RDF) and semantic web formats (RDF, Turtle).

How does GOSTAR compare to ChEMBL when it comes to building predictive models for drug discovery and development processes? Read on.

ChEMBL vs. GOSTAR – data diversity and compound coverage

Predictive modeling can be successfully used to assess a drug's absorption, distribution, metabolism, and excretion (ADME) profile. It requires the developed algorithms to be first trained on high-volume of highly diverse ADME data. Two of the most popular sources of data for building predictive models are GOSTAR and ChEMBL. Both are used by medicinal chemists, computational scientists, pharmacologists, and toxicologists to support drug discovery and development programs. The quality of their data is highly regarded in the pharmaceutical industry, and both GOSTAR and ChEMBL incorporate manual curation processes to maintain quality standards.

To minimize the risk of inaccurate prediction, the database on which the predictive model is trained must contain a large volume of highly diverse data. GOSTAR substantially exceeds the number of compounds, bioactivities, literature assets, and patents found in ChEMBL.

Database	**Compounds**	**Bioactivities**	**Scientific literature**	**Patents**
GOSTAR	9.4 million	32 million	208,901	90,614
ChEMBL	2.4 million	20 million	83,415	2,564

Table 1: Comparison of GOSTAR and ChEMBL database size

But an advantage in volume would be inconsequential if not matched by diversity. How does GOSTAR data contrast with ChEMBL data in this critical respect?

By using Konstanz Information Miner (KNIME), an open-source data analytics, reporting, and integration platform with tools and workflows for building machine learning and data mining models, we tested the number of unique compounds in GOSTAR vs. ChEMBL.

Our analysis established without doubt that GOSTAR contains more unique chemical structures for exemplified ADME parameters than ChEMBL.

The number of unique compounds in GOSTAR ranges between 2x and 7x that of those overlapping with the ChEMBL database.

ADME Parameter	No. of compounds in ChEMBL	No. of compounds in GOSTAR	Overlap*	Unique compounds in GOSTAR**	Mode
Caco-2 permeability	7,277	13,616	3,475	10,147	0.29
LogD	25,550	25,836	12,263	13,743	0.34
Madin-Darby canine kidney (MDCK) permeability	6,479	9,205	1,723	7,482	0.27
Plasma protein binding (PPB)	3,967	14,838	1,808	12,968	0.28
Human hepatocyte clearance	1,096	2,937	601	2,331	0.25
Rat hepatocyte clearance	1,129	2,763	629	2,135	0.24
Human liver microsomal clearance	9,252	14,819	4,388	10,478	0.32
Rat liver microsomal clearance	4,492	6,872	2,292	4,609	0.30

Table 2: Comparison of unique compounds in GOSTAR and ChEMBL. *Overlap = Number of compounds with fingerprint similarity of 1 between ChEMBL and GOSTAR. **Unique compounds in GOSTAR = <0.98 Tanimoto similarity

The implications of GOSTAR's clear advantage in diversity are profound:

Data scientists and computational chemists seeking greater accuracy from their predictive models are better served with ADME data from GOSTAR than from ChEMBL. How do we assure that GOSTAR is the largest and most diverse dataset?

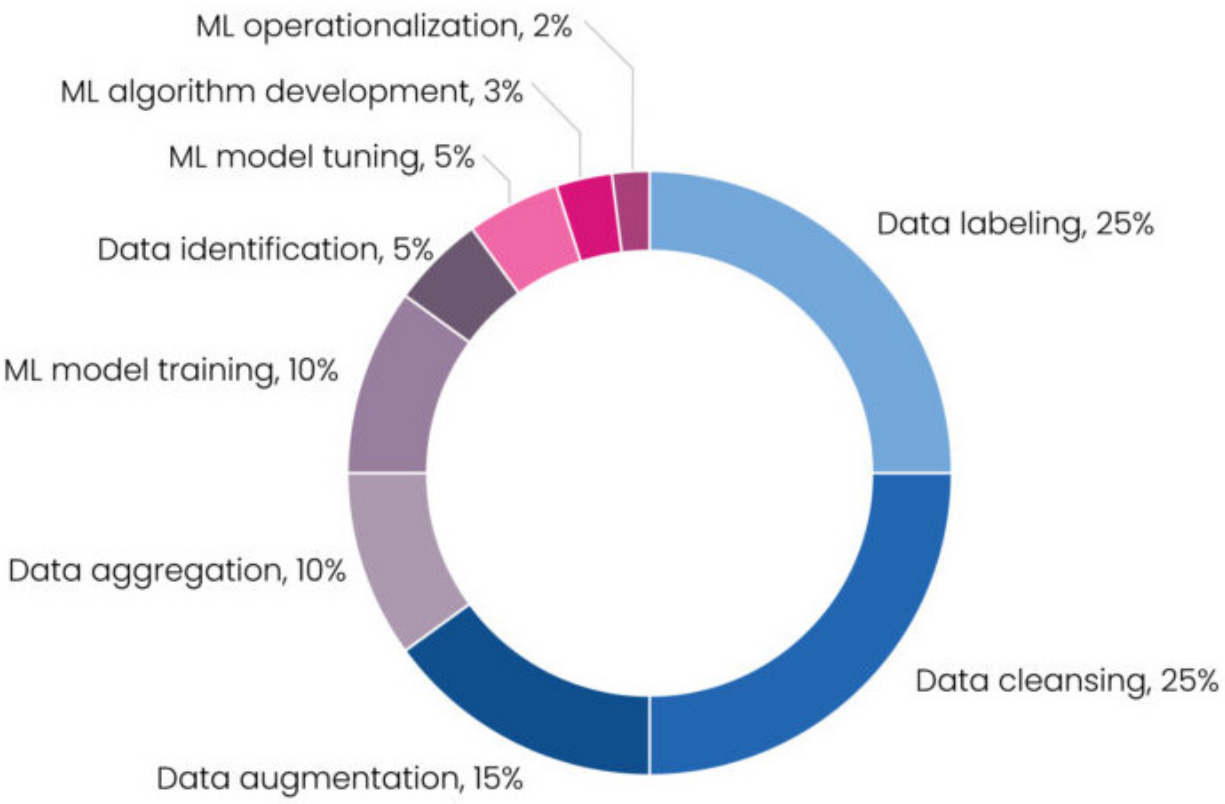
Excelra is the global leader in the manual extraction of SAR data from the scientific literature. We're regularly streamlining data collection and standardization processes for the world's leading pharma and biotech companies. This allows us to constantly expand our libraries so that more pharma and biotech companies can benefit from the time and cost savings we deliver.

See how thanks to our efficient data curation services a large US biopharmaceutical company engaged in KRAS research was able to swiftly proceed with its ongoing research program, avoiding the bottleneck of the data collection phase: KRAS program IP development using data curation.

Selecting and preparing data for AI/ML predictive modeling

High-quality data is the limiting factor in AI/ML predictive modeling

Despite the unquestionable benefits that ML modeling provides in the drug discovery and development process, data processing and cleaning are the bottlenecks and limiting factors in the ML approach. Approximately 80% of the time building an ML model is spent preparing data and only 20% writing the algorithm.



Data is the raw material on which AI/ML models are built and operated. But the quality of data is widely variable, and the heterogeneity of data sets – gathered from dozens of global sources at the expense of great resources – can be difficult to standardize and optimize for AI/ML analysis.

Choosing the best database for AI/ML predictive modeling

Above all, predictive accuracy relies on 3 key characteristics that databases must provide to facilitate high-performing models: quantity, quality, coverage

These three characteristics must be balanced to suit a model's specific task. Without a sufficiently high volume of data, a model's prediction is invalidated by its sample size. The qualifying criteria for a valid prediction will change depending on the type of prediction required. At the higher end, for example, an extremely large volume of data would be required for a model built to differentiate between active and inactive compounds across a highly diverse set of chemicals. At the other end of the spectrum, a lower number of data points are required for a refined quantitative model to optimize molecular interactions between a compound and its receptor as measured by X-ray crystallography. Yet, in both cases, data coverage is a big factor in determining valuable results. To greater or lesser extents, data of insufficient quality, polluted by meta noise and inconsistent formatting, could dramatically hinder the functionality of some models and lead to problematic analyses. So, in some cases more than others, data quality will be the principal consideration.

The quantity and quality of available data, and the breadth of its chemical coverage, become essential when selecting a database to build and train AI/ML models on. Excelra's GOSTAR is the world's largest database curated by domain experts, and its data quantity, quality, and coverage are exemplary. Comprehensively compiled and consistently structured, GOSTAR's data is also easy for a data scientist to prepare for a predictive model.

How to prepare a GOSTAR data set?

Step 1: Select the data

This step involves selecting a subset of all available data. Though there may be an urge to include all available data, it isn't always true that more data is better. More data is only necessarily better if it's relevant to the study's research question. Machine learning models can be misled if the training set is not representative of reality.

1. Retrieve all data associated with the target(s) of interest:
   1.1. Define the target(s) using Common_Name, , UniProt ID, or similar fields available in GOSTAR.
2. Ensure that data obtained from searching target(s) of interest in step 1 have chemical structures associated with the results:
   2.1. Chemical structures are required since the primary focus is on the structure-activity relationship (SAR) and AI/ML models relating chemical structures to bioactivities.

   2.2. Rows of data retrieved which do not have associated chemical structures can be discarded for this exercise.
3. Identify the experimental conditions:
   3.1. Ideally, data should be obtained from previous experiments conducted under similar experimental conditions to ensure a valid comparison.

   3.2. A conservative (and impractical) approach would be to only consider records that have identical experimental conditions. Including only these records would significantly reduce the data available for modeling.

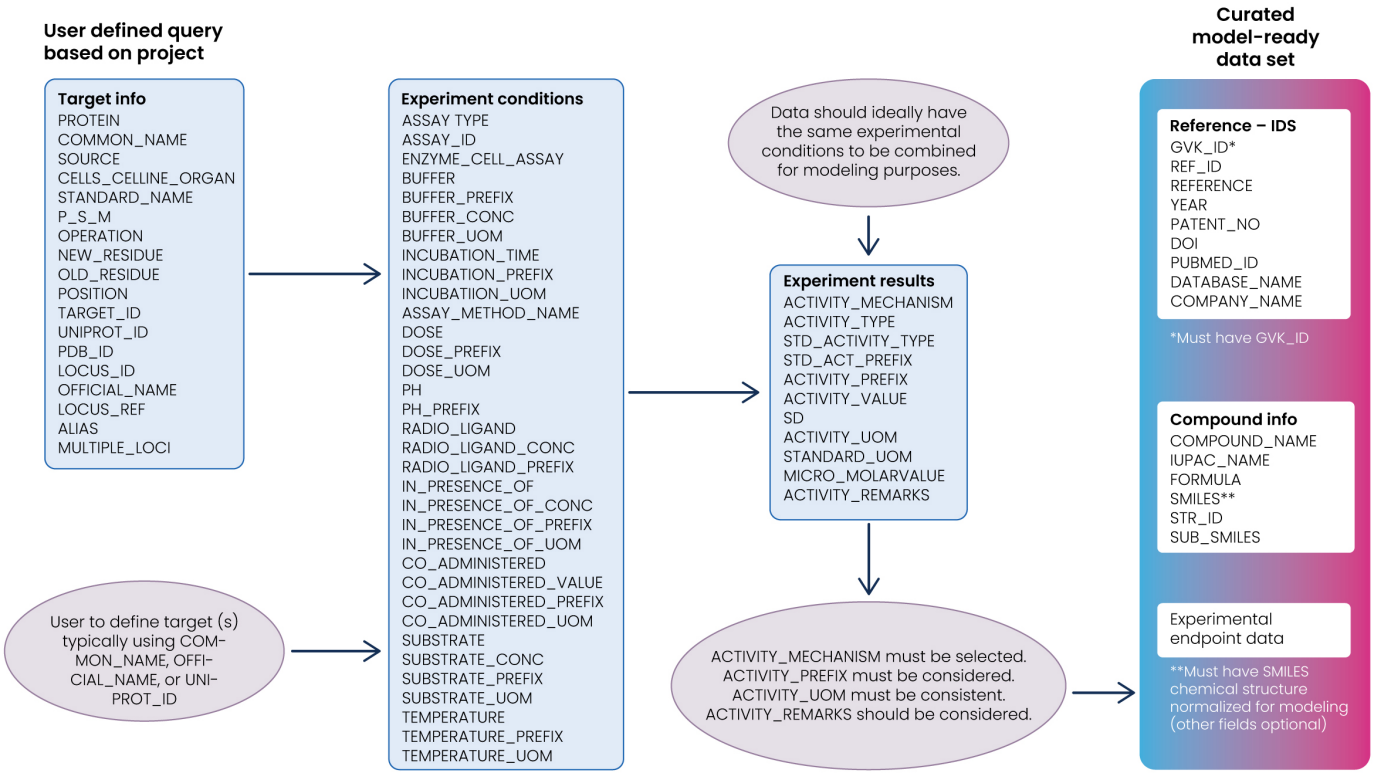
Consider the following questions:

1. How many of the experimental conditions' fields available in GOSTAR can be exact matches?

2. Which conditions are similar enough to allow for combining data from multiple studies?

3. How can we take advantage of studies which may not have explicitly recorded experimental conditions?

The answer to these questions is very subjective and requires agreement between data scientists and experimentalists. The experimental conditions in GOSTAR reflect the information available in publications and patents. GOSTAR enables rapid extraction of this information without having to read, analyze and curate data directly from hundreds or thousands of primary references.


User defined query based on project
Target info
PROTEIN
COMMON_NAME
SOURCE
CELLS_CELLINE_ORGAN
STANDARD_NAME
P_S_M
OPERATION
NEW_RESIDUE
OLD_RESIDUE
POSITION
TARGET_ID
UNIPROT_ID
PDB_ID
LOCUS_ID
OFFICIAL_NAME
LOCUS_REF
ALIAS
MULTIPLE_LOCI
User to define target (s) typically using COM-MON_NAME, OFFI-CIAL_NAME, or UNI-PROT_ID
Experiment conditions
ASSAY TYPE
ASSAY_ID
ENZYME_CELL_ASSAY
BUFFER
BUFFER_PREFIX
BUFFER_CONC
BUFFER_UOM
INCUBATION_TIME
INCUBATION_PREFIX
INCUBATIION_UOM
ASSAY_METHOD_NAME
DOSE
DOSE_PREFIX
DOSE_UOM
PH
PH_PREFIX
RADIO_LIGAND
RADIO_LIGAND_CONC
RADIO_LIGAND_PREFIX
IN_PRESENCE_OF
IN_PRESENCE_OF_CONC
IN_PRESENCE_OF_PREFIX
IN_PRESENCE_OF_UOM
CO_ADMINISTERED
CO_ADMINISTERED_VALUE
CO_ADMINISTERED_PREFIX
CO_ADMINISTERED_UOM
SUBSTRATE
SUBSTRATE_CONC
SUBSTRATE_PREFIX
SUBSTRATE_UOM
TEMPERATURE
TEMPERATURE_PREFIX
TEMPERATURE_UOM
Data should ideally have the same experimental conditions to be combined for modeling purposes.
Experiment results
ACTIVITY_MECHANISM
ACTIVITY_TYPE
STD_ACTIVITY_TYPE
STD_ACT_PREFIX
ACTIVITY_PREFIX
ACTIVITY_VALUE
SD
ACTIVITY_UOM
STANDARD_UOM
MICRO_MOLARVALUE
ACTIVITY_REMARKS
ACTIVITY_MECHANISM must be selected.
ACTIVITY_PREFIX must be considered.
ACTIVITY_UOM must be consistent.
ACTIVITY_REMARKS should be considered.
Curated model-ready data set
Reference – IDS
GVK_ID*
REF_ID
REFERENCE
YEAR
PATENT_NO
DOI
PUBMED_ID
DATABASE_NAME
COMPANY_NAME
*Must have GVK_ID
Compound info
COMPOUND_NAME
IUPAC_NAME
FORMULA
SMILES**
STR_ID
SUB_SMILES
Experimental endpoint data
**Must have SMILES chemical structure normalized for modeling (other fields optional)

Step 2: Pre-process and transform the data

The next step is to preprocess and transform the selected data. While the preprocessing step involves converting the selected data into a usable format, the transformation step is influenced by the algorithm used and understanding of the problem domain. Several iterative transformations to the processed data may be required, some of which are illustrated below.

Consideration of endpoint fields:

1. Inspect search results to determine how the majority of data is recorded: IC50, EC50, %Inhibition, etc. Combining observations from different endpoints is not usually reasonable unless rules are defined for qualitative modeling. Identifying the most prevalent endpoint will increase the quantity of useful data.

2. Activity prefixes must be considered and accounted for when combining results for modeling. Censured data, i.e., "<", ">" must be removed or accounted for in continuous models. Categorical models are less sensitive than continuous models to censured data.

3. Measurement needs to be consistent and converted to "standard units," which should be defined according to project needs. Only comparable units of measurement should be converted.

4. Activity remarks should be considered, as this field highlights potential inconsistencies documented in the primary sources.

General comments:

- The resulting data set must be carefully examined to remove rows that contain "null" or missing values if these values are critical to the modeling exercise.

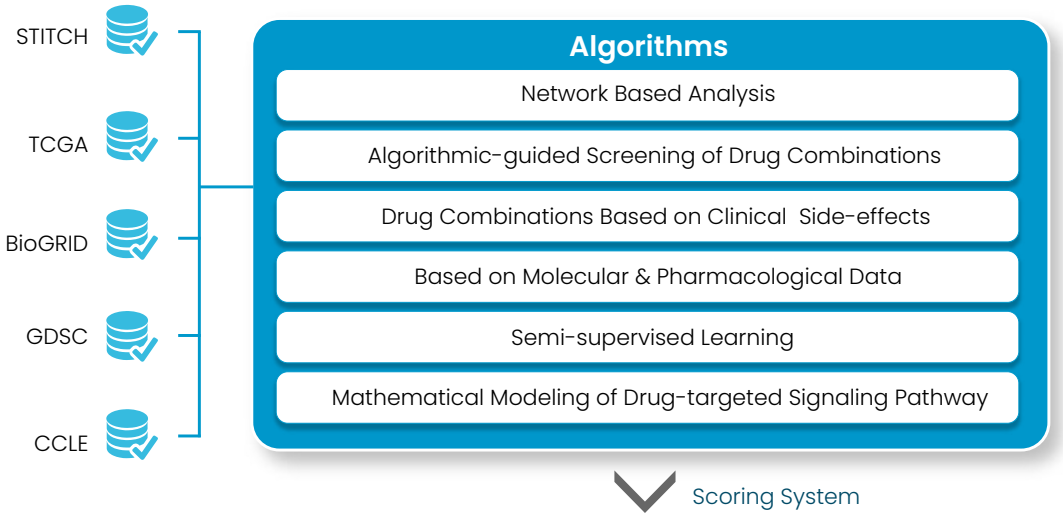
- Structures must be standardized and prepared for descriptor generation. There are numerous approaches to chemical structure standardization in the literature, which cover topics such as salt stripping, tautomer generation, etc.

- Extreme outliers of both the experimental endpoints (outside typical ranges of measurement) and chemical structures (polymers, mixtures, etc.) should be removed.

- Quantitative models (regression) may not necessarily be required depending on the ultimate use case, e.g., classifying compounds into bins according to predetermined criteria.

- Data scientists should consider the chemical diversity of the data set and estimate the probability of new compounds falling inside or outside the chemical space from which the model was derived.

- Clearly define the intended purpose of the model before embarking on the derivation of a model. Data selection should be adapted to meet the requirements.



Color scheme is based on Drug X response

	Overall cancer level				Each patient level insight			
	Anti-PD1 Sensitivity (RPART)	Anti-PD1 PLS score	Drug X PLS score	Both (Anti PD1+ DrugX) Sensitive	Total Sample	Sensitive sample (both Drug X+Anti PD1)	Resistant Sample	Response Rate (Anti-PD1 blocker)
ALL	0	2.08	0.55	Yes	750	131	619	17.47
LIHC	-1	0.01	0.31	Yes	373	185	188	49.60
PAAD	-1	0.74	0.29	Yes	179	57	122	31.84
OV	0	0.34	0.93	Yes	558	497	61	89.07
AML	0	2.26	0.72	Yes	542	77	465	14.21
BLCA	-1	2.44	-0.40	No	408	199	209	48.77
CHOL	**-1**	**-0.83**	**-0.30**	No	36	13	23	36.77
STAD	-1	0.41	0.61	No	415	164	251	39.52
TNBC BL1	-1	0.88	0.32	Yes	20	16	4	80.00
BRCA	-1	-0.20	-1.05	No				

RPART
0 → Resistant
1 → Sensitive
-1 → Sensitive

PLS score
Positive → Sensitive
Negative → Resistant

Yes → Must be Drug X positive and either of Anti PD1 predictor (RPART or PLS) as sensitive

Figure 4: The process of curating GOSTAR data for AI/ML applications

Conclusion

The democratization of artificial intelligence technology, combined with the introduction of big data and ever-increasing computing power, has accelerated AI/ML adoption, particularly in the pharmaceutical and biotech industries. However, the predictive accuracy of AI/ML models is dependent on the data that powers them. AI/ML advances are inhibited by data of insufficient quantity, quality, and coverage.

GOSTAR overcomes this problem by delivering comprehensive, homogenized, high-quality data typically missing from other SAR databases. Users can take full advantage of the millions of compounds and associated endpoints in the GOSTAR database by following a framework for selecting and preparing data. Once selected and prepared, GOSTAR data can help build optimized predictive models that could uncover major discoveries.

GOSTAR data alleviates bottlenecks, saves time, and improves accuracy. When technology depends on data, researchers depend on GOSTAR.

To learn more about GOSTAR and Excelra's data, insight, and R&D technology services, visit excelra.com/gostar.

Case study 2

Combination feasibility prediction for checkpoint inhibitors for a biologic

Client's challenge and goal

Our client, a biotech company based in Europe, had a large molecule in the development pipeline for cancer indications. The client was interested in combining their proprietary molecule with already approved immune checkpoint inhibitors to improve therapeutic efficacy.

The client requested to prioritize cancer indications based on their sensitivity towards the combination of the biologic with a checkpoint inhibitor (anti-PD-1/PDL-1).

Our approach

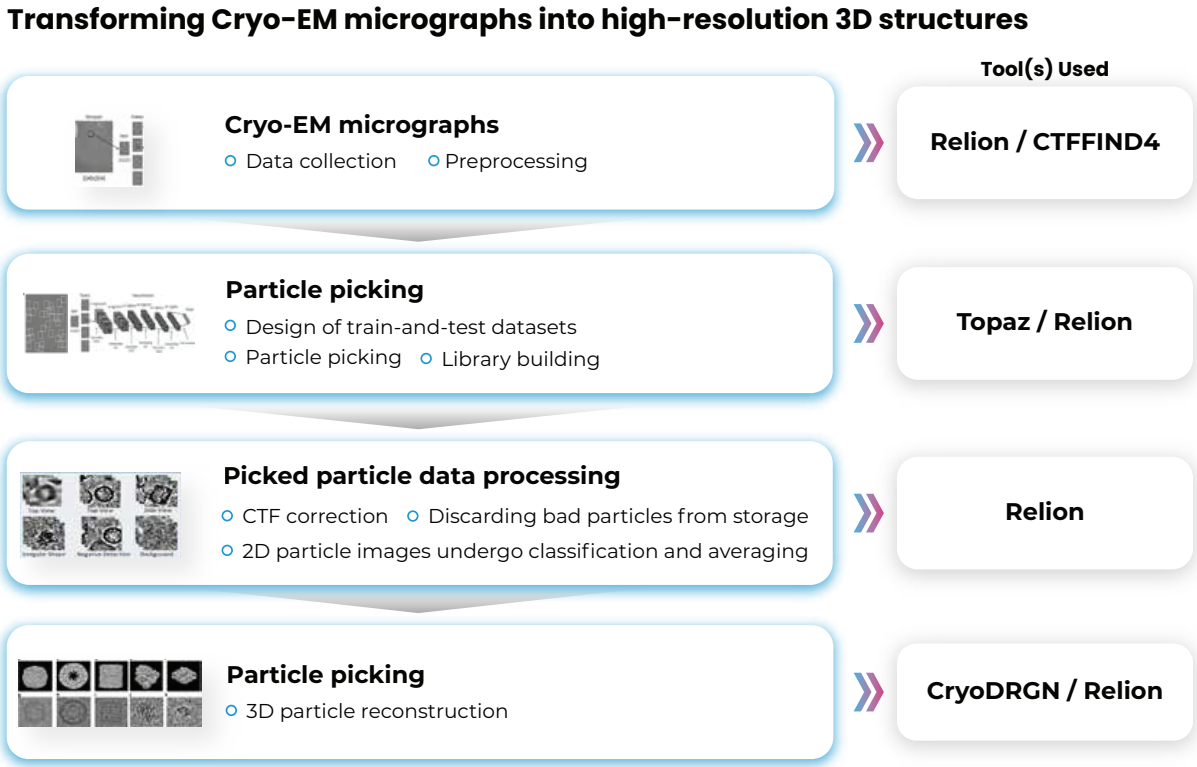


Figure 5: Cryo-EM data analysis workflow in Excelra's pipeline

For building predictive models, the Excelra team used publicly available data on successful and fail ed drug combinations.

We built Machine Learning models to assess the sensitivity of cancer indications as well as patients to the drug combination. Based on the analysis, some cancer indications were prioritized for further assessment. We then developed a biological hypothesis to establish the synergistic role of the combination partners for cancer treatment.

Result and impact

With our results, the client was able to prioritize the indication where therapy with PD-1 will work the best.

“Indications resistant or were partially sensitive to the monotherapy were predicted to be sensitive towards combination with the checkpoint inhibitor.”

The client received moreover a widened list of indications where the query drug may be developed and the feasibility/synergy prediction of the two-drug combination.

We also generated custom pathways to uncover crosstalk between the drug-induced signaling and checkpoint inhibitor signaling pathways.

Case study 3

Powering up cryo-EM data for faster drug discovery

Client's challenge and goal

One of our large pharma clients sought to improve their internal drug discovery program and derive value from their OMICS data by establishing an enterprise-wide R&D-IT ecosystem specific to the early discovery and translational function.

Requiring experts who can integrate scientific knowledge with technological innovation, the company approached Excelra. We were tasked with developing an end-to-end pipeline for processing cryogenic electron microscopy (Cryo-EM) images to learn more about protein structures and their conformational and compositional heterogeneity.

Our approach

Cryo-EM is a highly specialized technique, but our team had significant experience within the field and clearly understood the requirements. The technique produces 2D maps containing several particles. These maps are low resolution, randomly oriented, and intrinsically heterogeneous, making analysis difficult.

To overcome this difficulty, we designed a pipeline incorporating cutting-edge AI/ML technologies to transform the original micrographs into high-resolution 3D structures.

Result and impact

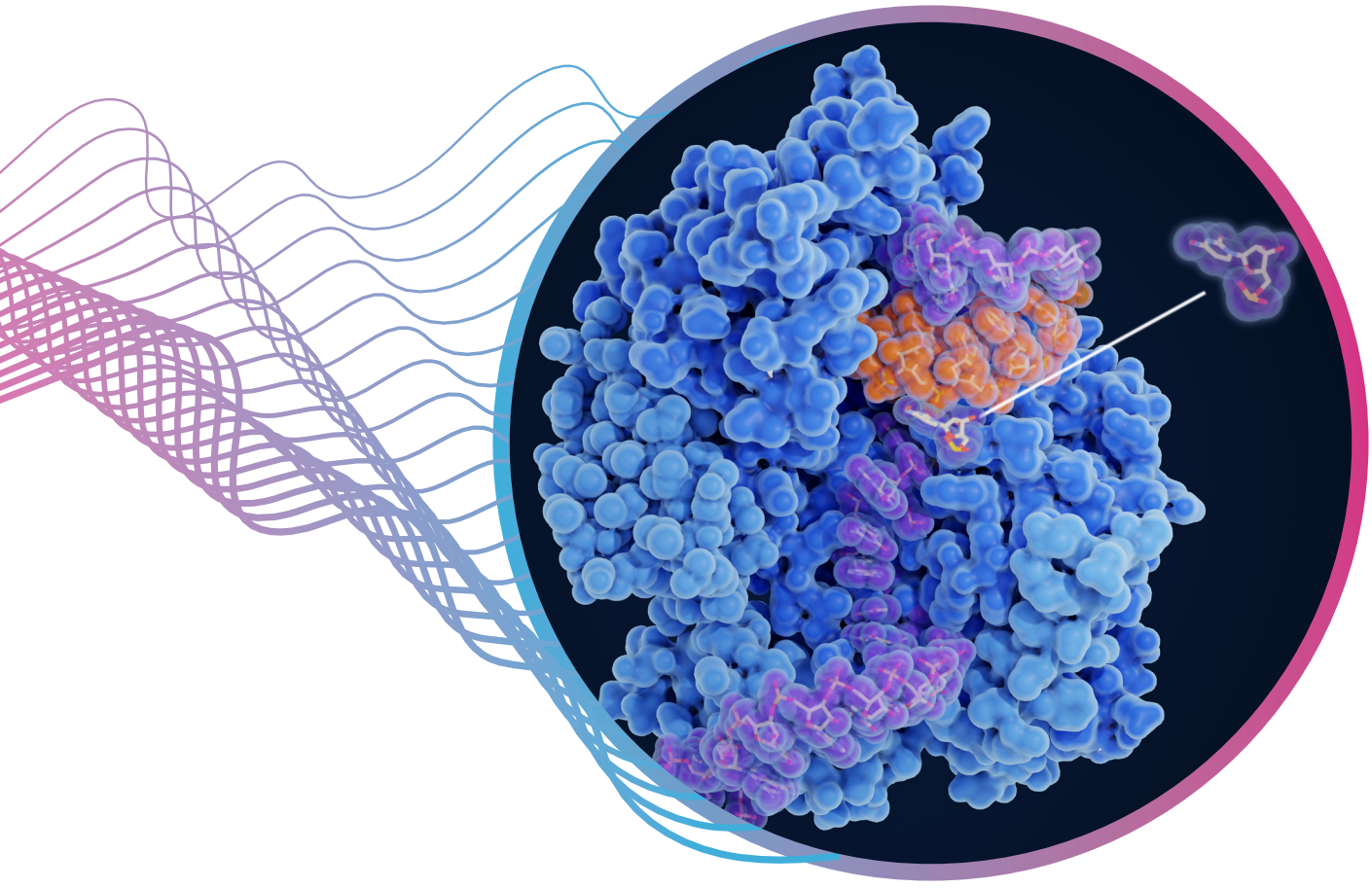
The successful deployment of the pipeline has dramatically saved time, reduced cost, and improved outcomes for protein target credentialling. It has improved the analysis of structures that are essential in therapy development.

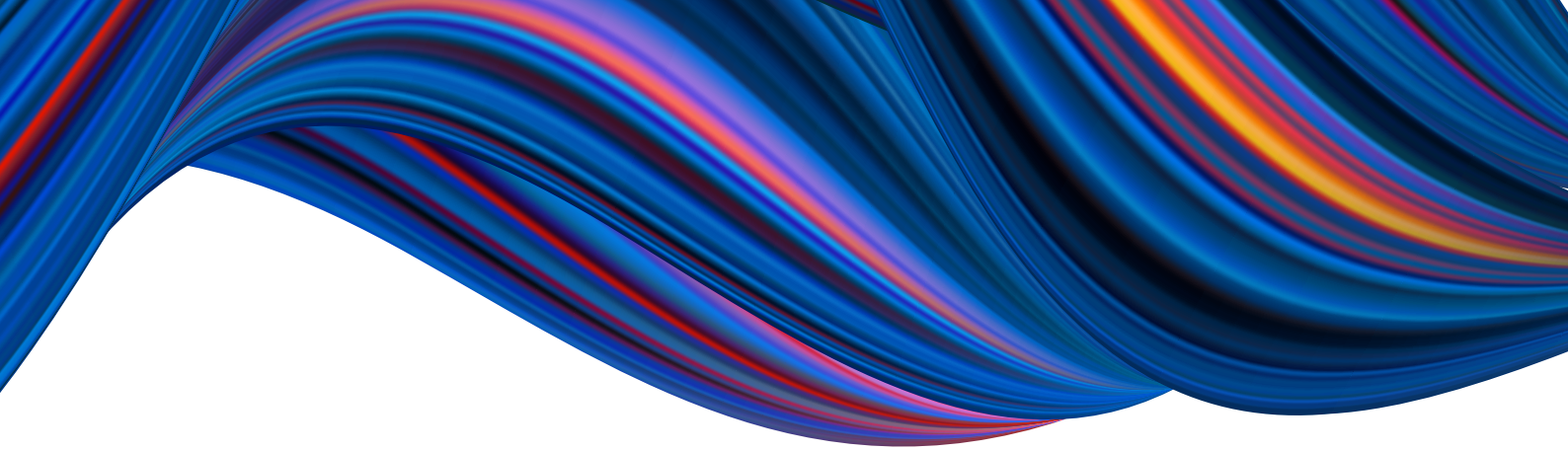
The client was able to accelerate their discovery journey by obtaining from our team:

- Accurately annotated data, curated to avoid pollution or irrelevance and transformed to become usable within their existing systems
- A powerful sample workflow management system that's now in use across the whole organization, integrated with the cryo-EM pipeline and routinely benchmarked and tested to ensure impeccable performance
- Configuration of job schedulers like Apache Airflow, NextFlow, and SnakeMake, for batch processing on Azure, AWS, and GCP
- Biomedical search and cataloging tools that expedite the retrieval of internal studies enhanced with public data

- Automation of ETL and AI/ML pipelines, and development and deployment of OMICS data analysis pipelines on platforms like DNANexus and Seven Bridges
- Semantic data modeling that's transformed legacy datasets from donor samples and clinical trials.

Our ability to combine deep scientific knowledge and cutting-edge technology impressed the customer, and we now have a team of 70 experts embedded with them to support their objectives.





Chapter 5

Pharmacogenomics and biomarker strategies

Bioinformatics allows researchers to study the genetic basis of individual drug responses, an essential aspect of personalized medicine. By analyzing genetic variations among patients, bioinformatics can help identify biomarkers that predict drug efficacy and adverse effects, thus enabling the development of more effective and safer therapies tailored to individual patients.

Case studies

1. Identification of predictive biomarkers and applications in patient enrichment strategies
2. Bioinformatics and machine learning for biomarker discovery using public epigenetics data
3. Biomarker discovery in tropical disease

Case study 1

Identification of predictive biomarkers and applications in patient enrichment strategies

Client's challenge and goal

Our client, a small pharma company based in the United States, had a pipeline molecule that was under clinical development. The client was interested in identifying biomarkers indicative of drug-response in patients and further utilize the biomarkers for patient stratification in clinical trials.

The goal was to get actionable insights from proprietary gene-expression data of 118 cell lines that were treated with the drug. After the prediction of drug-response biomarkers, the client shared moreover gene expression profiles of 11 patients and requested to retrospectively classify them into responders and non-responders.

Our approach

Our team built machine learning models using three different methods to prioritize biomarkers associated with drug response. We performed pathway enrichment analysis to understand the role of the biomarkers in disease pathophysiology. Stratification of patients based on these biomarkers resulted in correct prediction of drug response in 8 out of 11 patients.

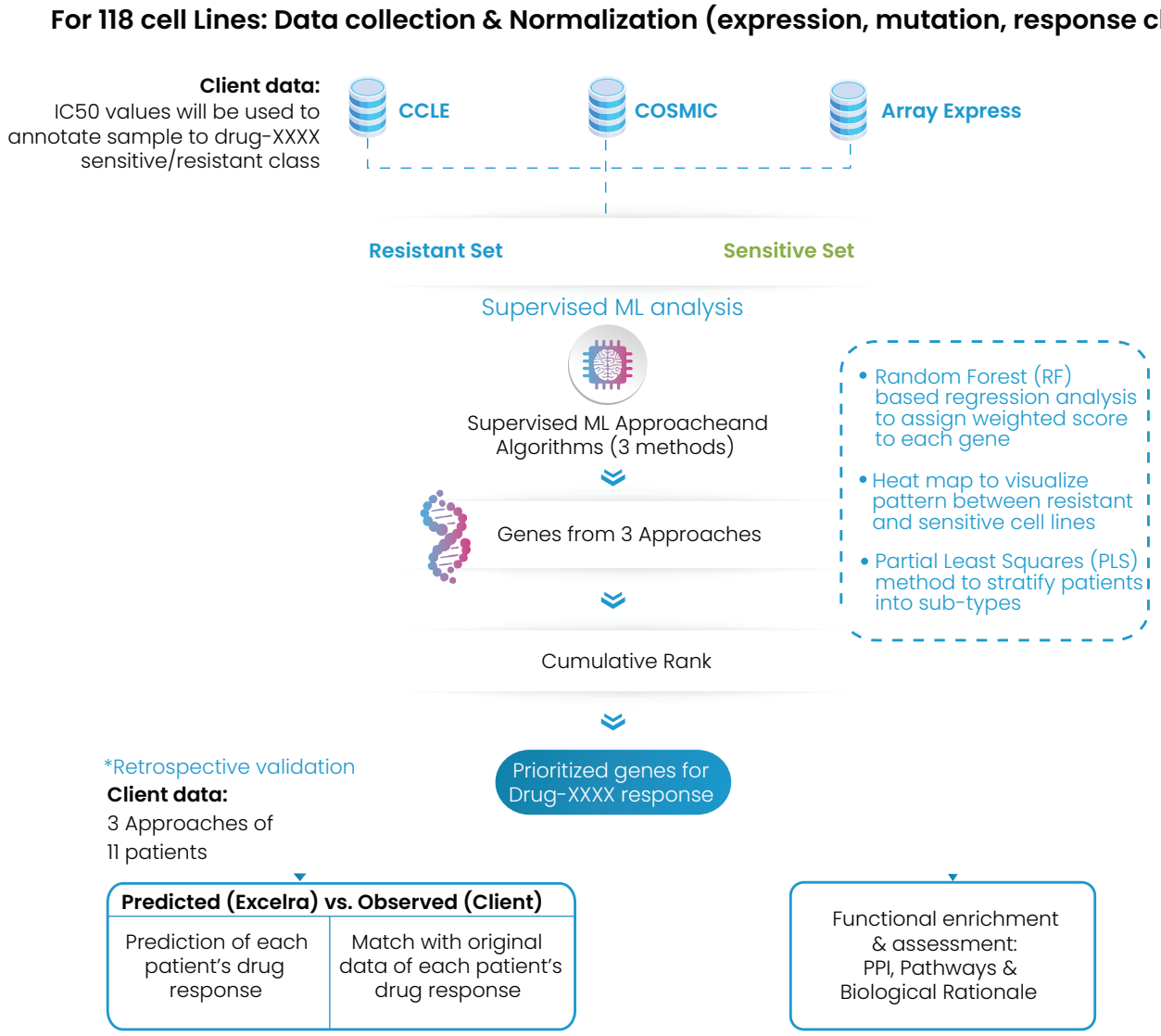


Figure 1: Our workflow in data collection and normalization (expression, mutation, response class) for the 118 cell-lines that were treated with the drug.

Result and impact

The client received 8 biomarkers identified for drug response. We predicted correctly 9 out of 11 patients' data yielding 82% prediction accuracy.

Our results enabled the client to transition from NHL to other tumor types.

Further deliverables:

- Gene signatures used to perform sub type-level analysis and patient stratification
- Establish the immune-modulatory role and defined Mechanism of Action (MoA)
- Opened possibilities for combinations with Immuno-Oncology agents

Case study 2

Bioinformatics and machine learning for biomarker discovery using public epigenetics data

Client's challenge and goal

Our client wanted to use public data to find biomarkers for several types of cancer. These biomarkers could then be included on the microfluidics assays they were developing.

Our approach

Our team helped the client to find and analyze relevant public datasets from various sources. We used our expertise in bioinformatics and machine learning to determine optimal epigenetics biomarker panels for the different cancer types of interest.

Result and impact

Thanks to our help, the client was able to effectively reduce their discovery and clinical trial costs.

Case study 3
Biomarker discovery in tropical disease

Client's challenge and goal

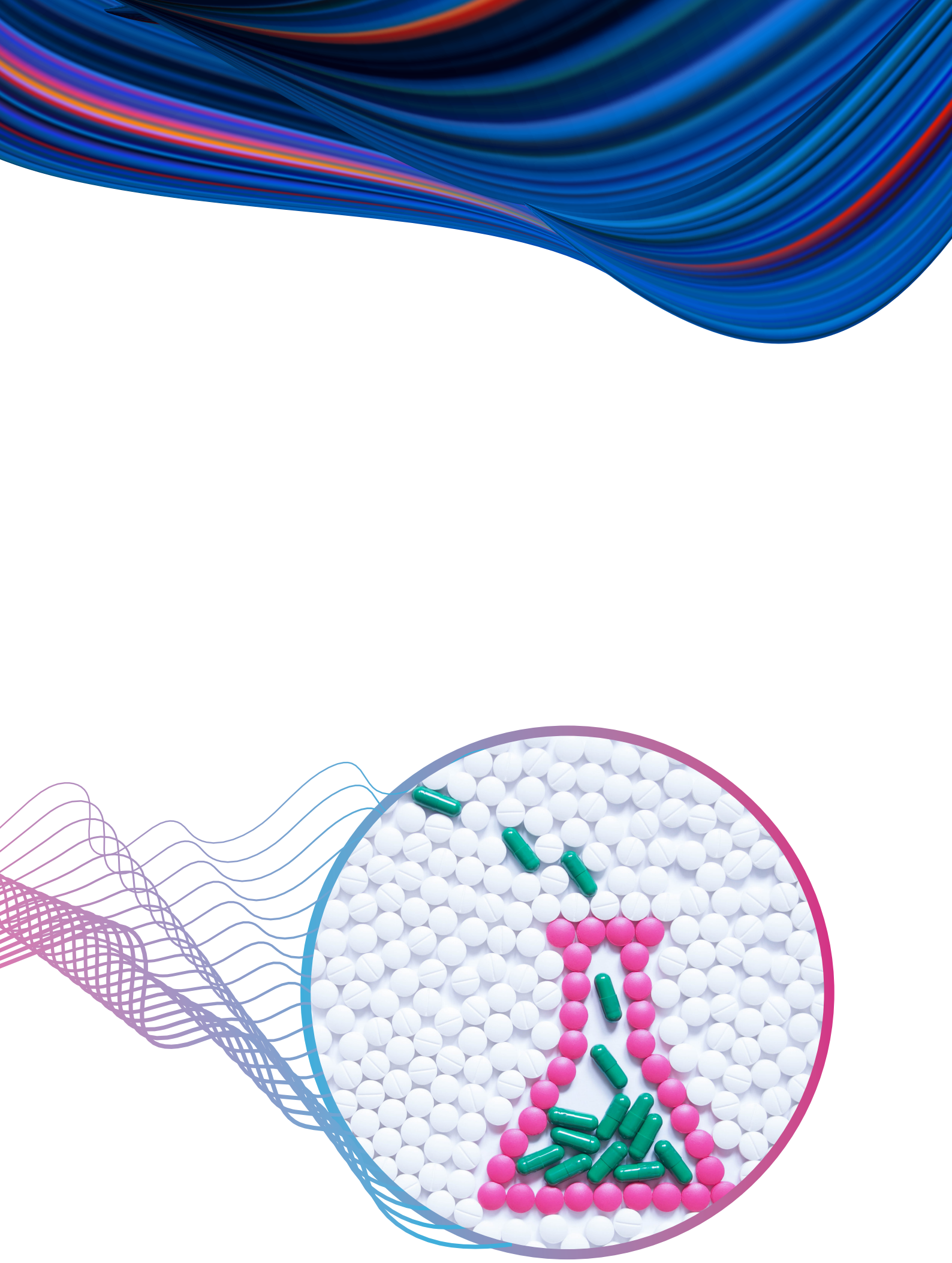
Our EU-based client was interested in performing a biomarker discovery study for a number of tropical diseases. They decided to have a large custom peptide array developed and then run patient vs control samples to determine which epitopes on which proteins could be used to develop diagnostic assays.

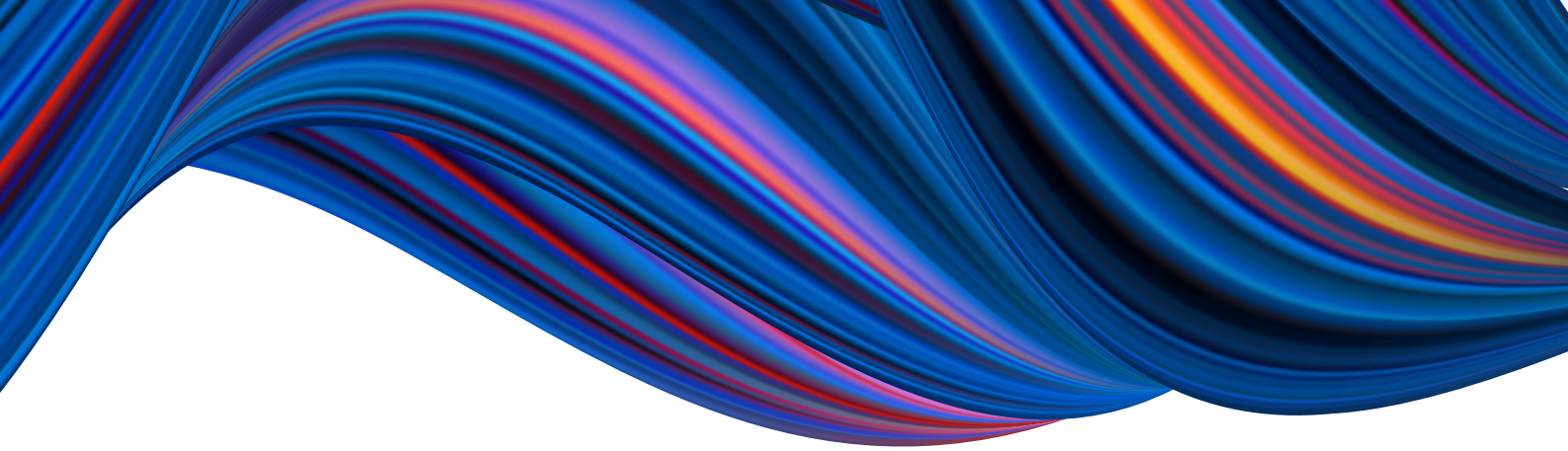
Our approach

Our team has helped design the peptide array based on the existing proteomes for Dengue, Yellow Fever, etc. Once the peptide array studies were finished, we built a custom analysis pipeline and helped the client with their analysis and determine possible relevant biomarkers.

Result and impact

With our support, the client was able to obtain the peptide array based on the existing proteomes for Dengue, Yellow Fever, etc. Once the peptide array studies were finished, our built a custom analysis pipeline and helped the client with their analysis and determine possible relevant biomarkers.





Chapter 6

Drug repurposing

We help identify new uses for existing drugs by analyzing large-scale data on drug-target interactions, gene expression profiles, and protein-protein interactions. This approach can significantly reduce the time, effort, and financial commitments required for traditional drug discovery methods.

1. Accelerated drug repurposing using advanced analytics

Case studies

1. Identifying 30 alternate indications for six shelved compounds
2. Identification of alternate indications for a clinical compound using ML

Case study 1

Identifying 30 alternate indications for six shelved compounds

Client's challenge and goal

A large US-based pharma company was looking to expand its portfolio for six shelved assets using drug repurposing. Four of the shelved assets failed in clinical development Phase II and III and two at the preclinical stage.

The client was interested in identifying alternate indications for their assets by an integrated strategy to get data that will inform which indication should be prioritized. The required data were obtained through a combination of in silico analysis with Excelra's proprietary platform and validation in wet lab experiments.

Our approach

We analyzed the six compounds with a comprehensive in silico pipeline that included Excelra's proprietary repurposing platform, 'GRIP' to generate a biological rationale for its use in new indications. Further, we conducted validation using a Multi-OMICS approach, performed in a wet lab to provide a final list of indications per compound.

Solution strategy


In silico analysis


Omics analysis


Correlation of both in silico and omics analysis

Overview of In-silico analysis

Excelra's novel and proprietary repurposing platform Global Repurposing Integrated Platform (GRIP) allows for an integrated approach by leveraging the following core components:

GRIP

The customized Global Repurposing Integrated Database is a compendium of 40+ public and proprietary databases and creates multi-dimensional profiles of biologically relevant entities such as genes, pathways, biomarkers, and adverse events

The holy trinity

3-way relationships established between drug, disease, and target help understand relevant associations

Algorithms

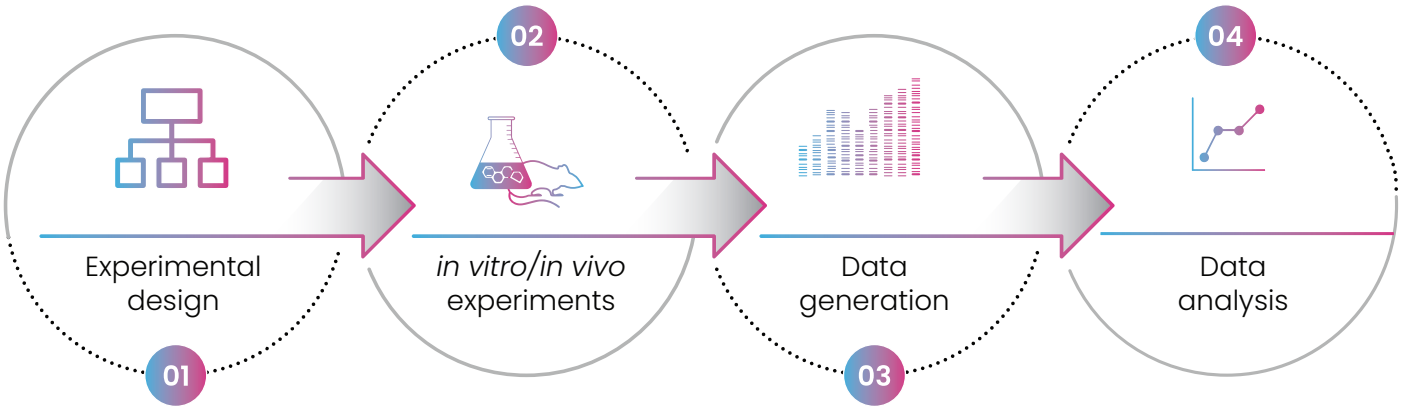
Customized proprietary algorithms have been developed internally to mine the drug-target-disease associations from GRIP

Analytics and visualization

In-house analytical capabilities and visualization tools enable the identification of hitherto hidden connections between drugs, diseases, and targets

Overview of omics analysis

Excelra's capabilities in handling OMICs data range across multiple data archetypes. In this engagement, the clients shared drug-treated in vitro and in vivo genomics and metabolomics data for the asset based on the recommendations from the previous in silico analysis. The multi-OMICs analysis approach used here consisted of analyzing that data to further validate the shortlisted alternate indications that could be recommended to the clients. The results from the multi-OMICs analysis were then used to corroborate the in-silico analysis to state the final recommendations, with all the supporting biological rationale in place to validate the outcomes.



Result and impact

The client was able to expand its portfolio by leveraging data-driven (in silico analyses) approaches corroborated with wet lab validation (multi-OMICs analyses). Ultimately, the findings enabled robust recommendations.

Accelerated drug repurposing using advanced analytics

Expertly curated data and advanced analytics have had a huge impact on drug repurposing. The importance of drug repurposing cannot be overstated, and significant advances in patient outcomes are achieved thanks to this branch of pharmaceutical science.

Whether researchers use a drug-centric, target-centric, or disease-centric approach to drug repurposing, the use of precisely selected and effectively analyzed data has been fundamental to the rapid growth in the number of repurposed drugs tested in the clinic.

Drug repurposing is powered up by machine learning

Advanced machine learning and data-driven analytics are having a profound impact on drug repurposing. The breadth and depth of the data produced by biological experiments have increased dramatically, allowing modern analytical tools to interrogate and interpret it in many different ways.

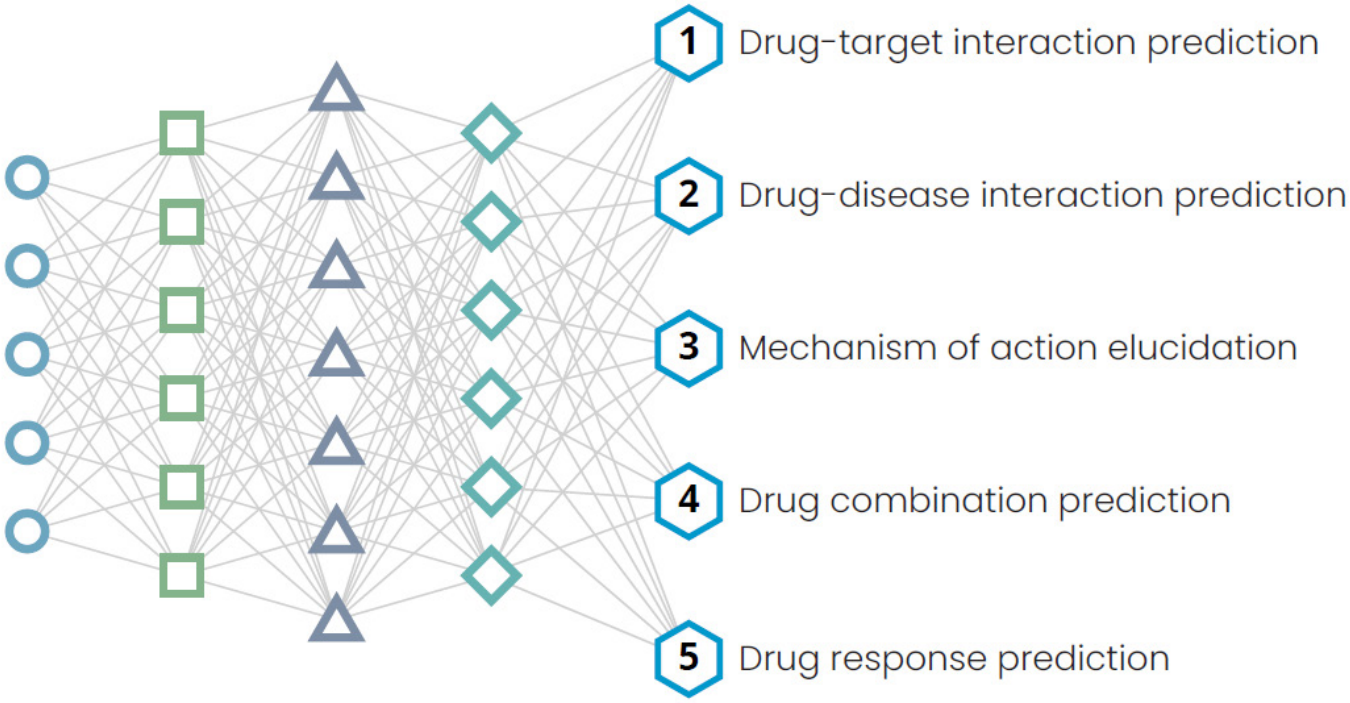


Figure 3: Application of machine learning repurposing

Excelra's role in accelerated drug repurposing

Excelra leverages our custom-built Global Repurposing Integrated Platform (GRIP) algorithms to mine public and proprietary databases to curate high-quality data for use in advanced analytics and predictive ML model building. Our experts work at the intersection of science and technology, and are uniquely adept at bridging the gap between those two fields.

Excelra is one of the most enthusiastic proponents of the data revolution and is a passionate advocate of utilizing advancing technologies in data analytics for drug repurposing programs. Data mining, structuring, transformation, and predictive model building are the core of our expertise.

Combining subject matter expertise with innovative technological solutions, we provide deep biological insights and powerful data outputs to provide exciting, value-driven recommendations to our clients.

Case study 2

Identification of alternate indications for a clinical compound using ML

Client's challenge and goal

A pharmaceutical client from the United States intended to repurpose their clinical candidate to identify alternative indications.

Our approach

Using data extracted by Excelra's GRIP platform, alternate indications were identified for the company's clinical compound. The pharma was also able to predict drug-disease associations by processing the data with advanced machine learning models

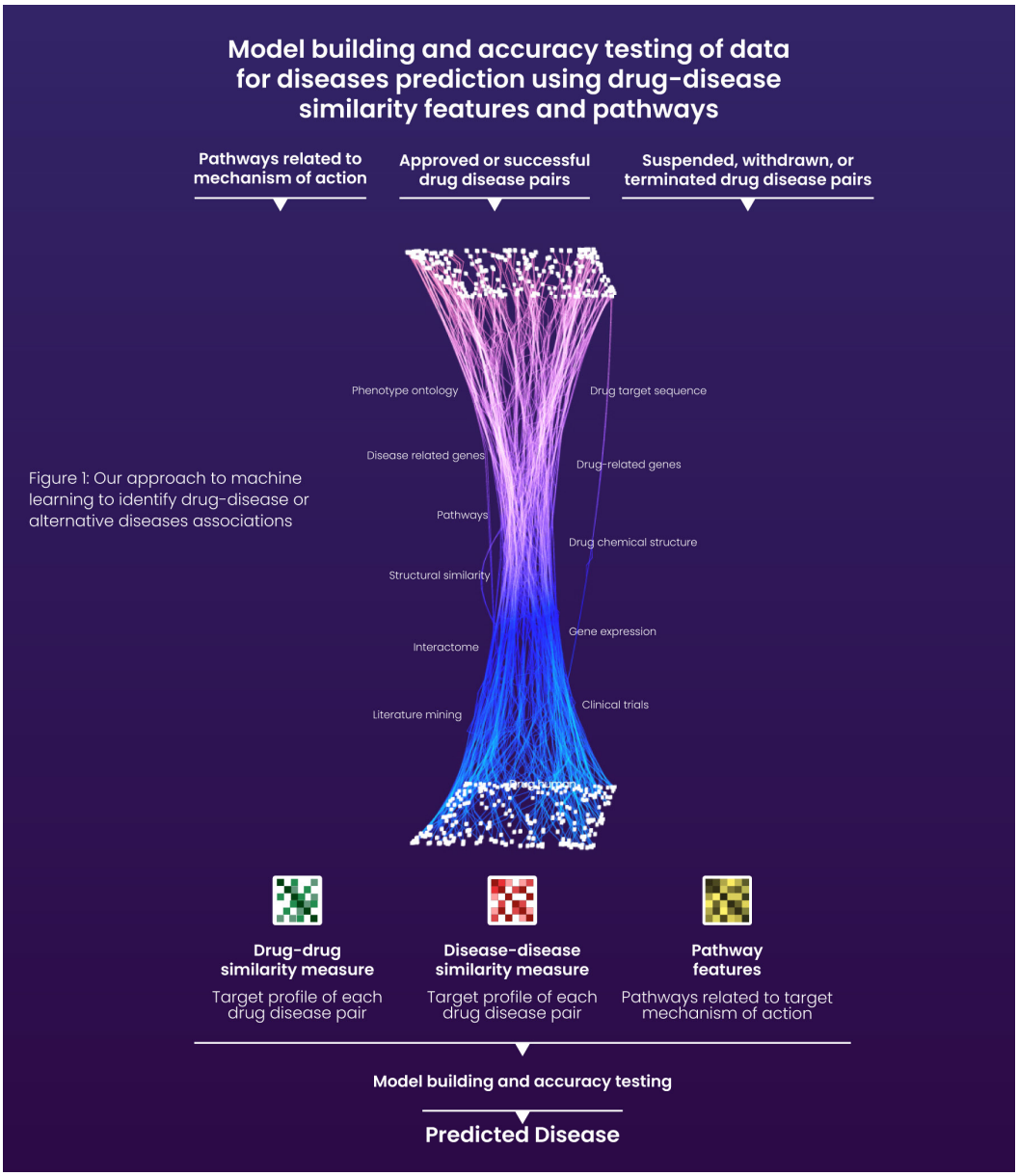


Figure 1: Our approach to machine learning to identify drug-disease or alternative diseases associations

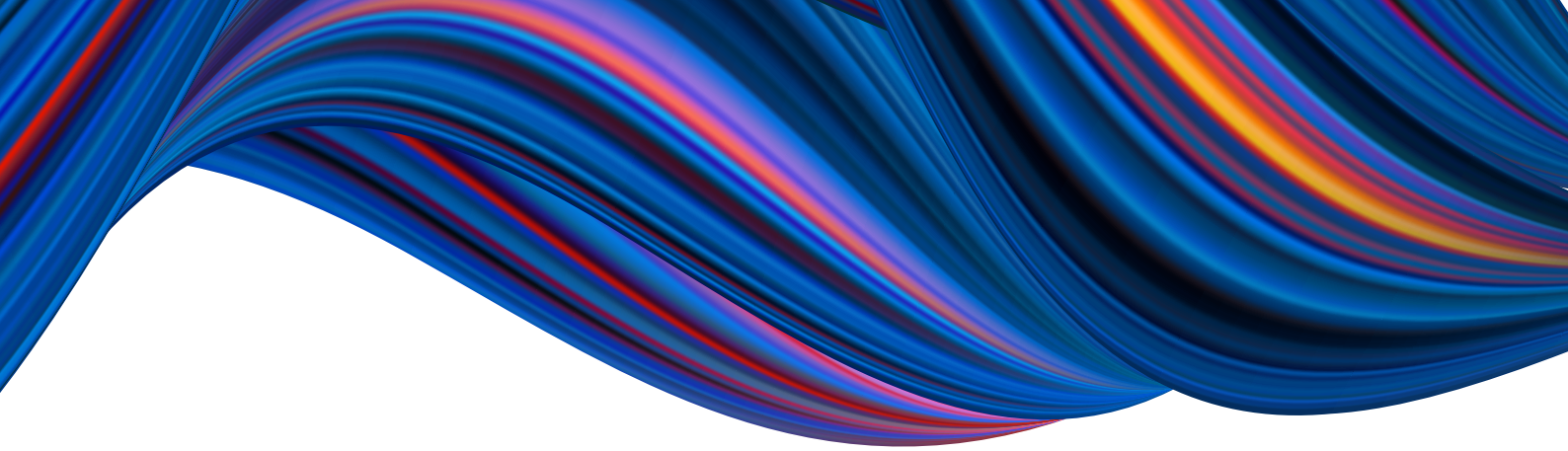
Figure 4: Excelra's approach to machine learning to identify drug-disease or alternative indications associations

Excelra's GRIP platform incorporated in silico analyses with literature mining, clinical trials, structural similarity, and interactome. We employed various machine learning approaches to identify potential drug-disease pairs, including drug-drug similarity, disease-disease similarity, and target mechanism of action-related pathway features. These approaches helped to predict several relevant drug-disease associations beyond those reported in the public domain for the scrutiny of subject matter experts.

Result and impact

The process returned recommendations for five drug-disease associations, supported by scientific rationale. The associations included two rare diseases, an autoimmune disease, a gastrointestinal disease, and a cardiovascular disease.

Excelra's curated data was used by advanced ML models to assess alternative drug-disease pairs. The astute application of data led to the identification of novel indications for the customer's clinical compound from diverse therapeutic areas and created significant additional value within the program.



Chapter 7

R&D informatics

The integration of informatics in drug discovery and development can significantly improve efficiency and reduce costs. In this chapter, we will explore how innovative solutions and technologies can facilitate the migration from legacy systems to modern, scalable platforms, as well as how collaborative workflow management systems can optimize processes and communications. We will also highlight the benefits of cloud environments and automation in enabling faster and more effective discovery of novel therapies, providing a competitive edge in the field of drug development.

Case studies

1. Cloud migration of R&D compound platform
2. Developing user-friendly and future-proof workflow management
3. Cloud implementation for co-development of automated ML analysis for target discovery
4. A robust and high-throughput pipeline for immune repertoire data analysis

Case study 1

Cloud migration of R&D compound platform

Client's challenge and goal

The client is a drug discovery company focused on small molecule oncology and immunology. A recurring phase of its R&D activity was the registering of compounds on a local database. The compound database was a desktop application that had been developed and primarily operated by a single user. The application was built on Microsoft Excel with unoptimized code, minimal functionality, a clunky UI, and limited security safeguards.',

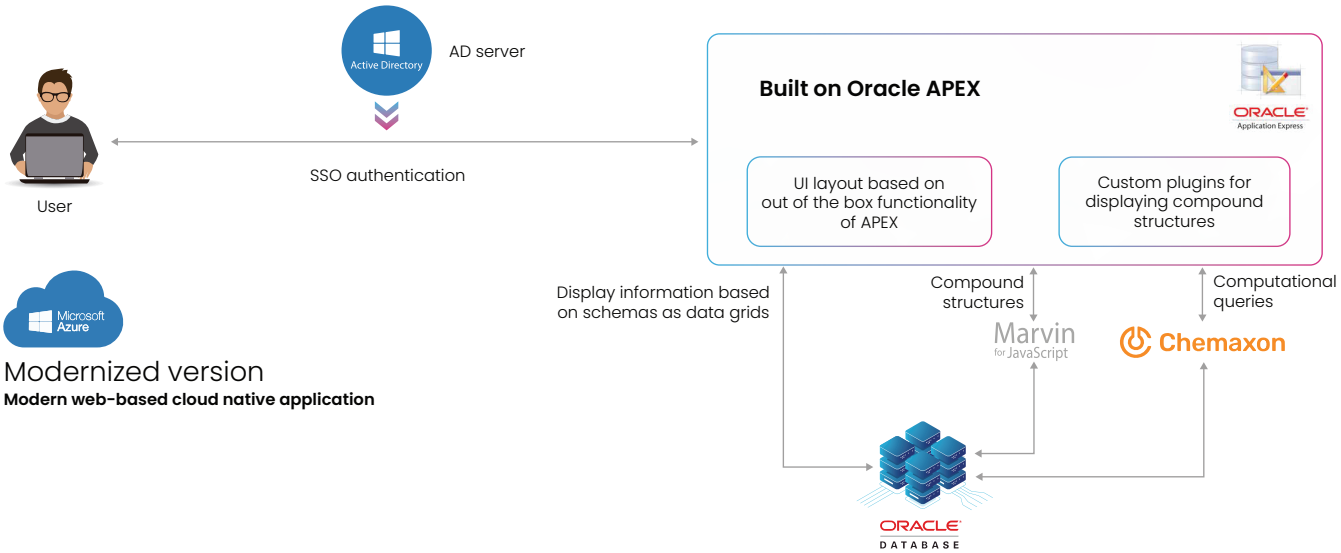
The client had two core requirements:

1. Replace the legacy database with a modernized, cloud-based web application.
2. Add a suite of features to improve search capabilities, import and export functions, compound comparison, remote working, team collaboration, and user access management.

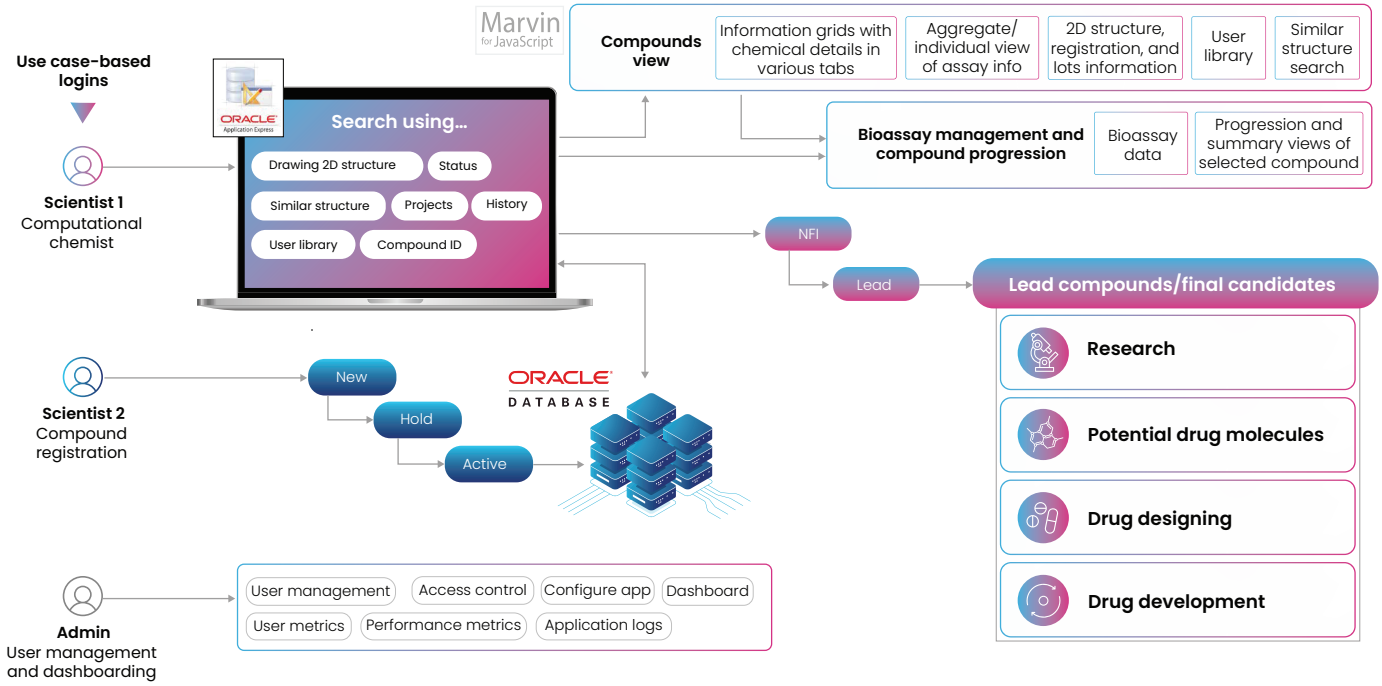
Our approach

After evaluating the existing application and the functional requirements within the wider workflow, our product design team worked with the client to shape the solution to its demands and presented a comprehensive proposal. Given that data was held on Oracle, an alternative was proposed to build the new application in Oracle APEX–a more flexible option with low-code development functionality baked-in.

Project overview



Platform capabilities



Result and impact

With our help, the client was able to successfully migrate from a slow, functionally limited, single-user, legacy platform to a modern, functionally robust web application that could be used at scale across the whole enterprise. The migration was executed without any loss of service or security compromises, and the client's team was given sufficient tools and training to quickly onboard users to the responsive, intuitive platform. The approach we proposed provided significant time and cost savings during development. It will continue to improve efficiencies and reduce expenditure thanks to the technologies utilized and the design methodology of our product development and cloud enablement teams.

The success of this project highlighted the efficacy of our approach. By combining domain expertise with technological innovation, we provided an exceptional solution that incorporated the flexibility required in drug discovery R&D without compromising on technical functionality.

Case study 2

Developing user-friendly and future-proof workflow management

Client's challenge and goal

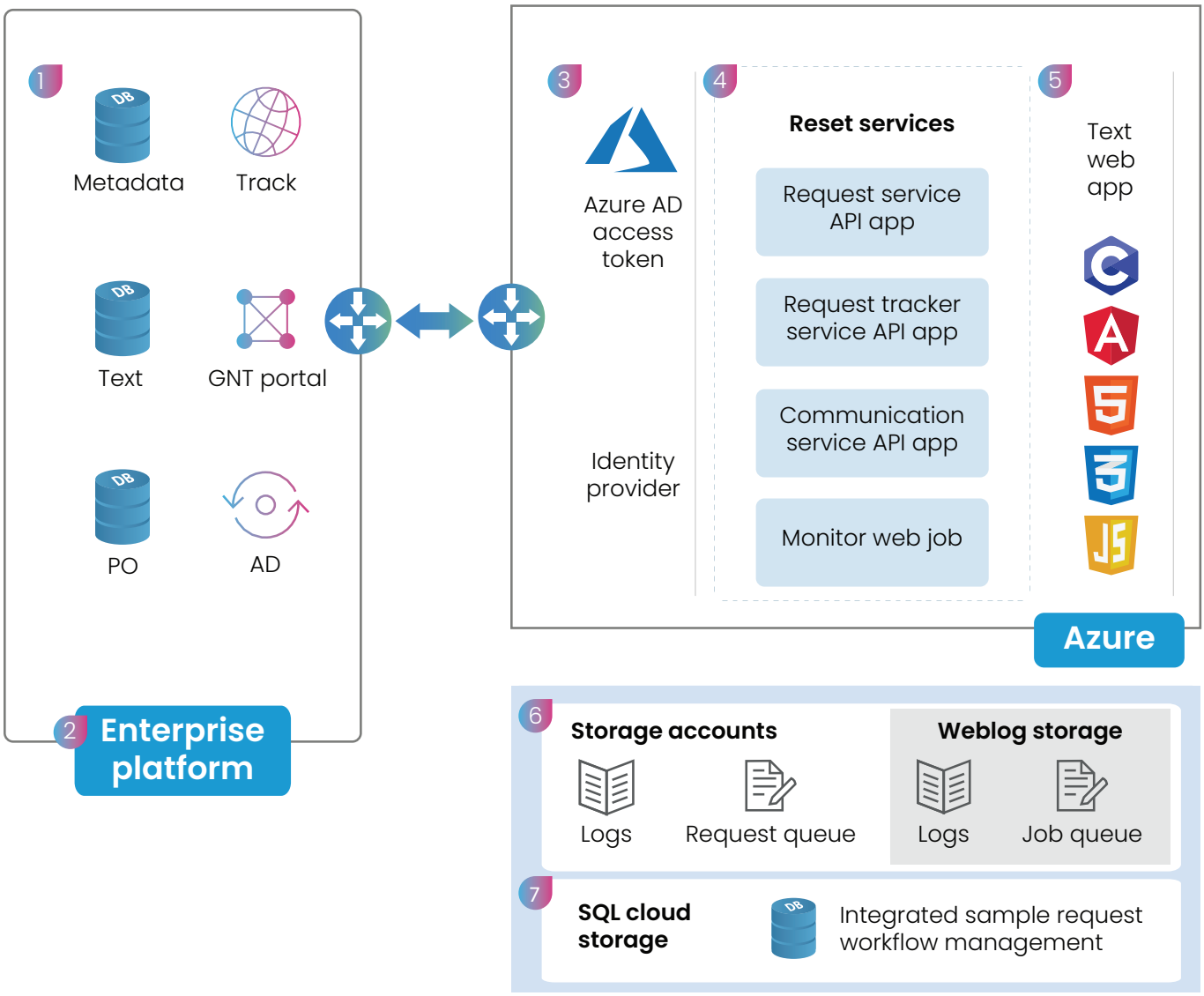
A life sciences company, based out of the US, partnered with Excelra to develop an organization-wide Collaborative Workflow Management System with a simplified user interface for internal and external stakeholders.

This system, while being user-friendly, had to be future-proof with automation/RPA-free, modular as well as scalable while being fluid and integrated among the various stakeholder systems across the ecosystem.

Our approach

An Azure Platform was designed on a server-less architecture. The platform had a simplified interface with integrated and collaborative workflow, metadata strategy, data lineage, and master data integration along with audit and compliance requirements, cross-collaboration between multiple CROs, and an inbuilt standardized vocabulary to align across datasets.

Architecture



Result and impact

Our client's processes and communications benefit from the Collaborative Workflow Management System that provides:

- Dynamic UI (admin controls) for front-end and back-end activities
- Creation, tracking, and validation of a single source for metadata
- User-friendly interface with all the data entry points in Excel format
- Scalability, Dynamic Framework, and Modular based add-ons
- Fluidity and departmental integration points across the framework
- Automation-future RPA Free

Case study 3

Cloud implementation for co-development of automated ML analysis for target discovery

Client's challenge and goal

Our client wanted to implement machine learning on their data, for which external input was required. However, the client didn't have the right infrastructure for the demanding computing actions. The goal was to build a suitable cloud environment.

Our approach

Our team built a custom and optimized cloud system according to the following requirements:

- Regional Set up f.e. GDPR compliance for EU clients
- Secure Cloud Environment with private subnets
- Optional security: MFA, NACLS
- Versioning via shared Github
- Direct access to a cloud server for data scientists
- Direct access to data via shared FTP/S3 download links

Result and impact

The customer saved significant costs thanks to the highly effective cloud environment created by our team.

Case study 4

A robust and high-throughput pipeline for immune repertoire data analysis

Client's challenge and goal

Our client needed a robust and high-throughput pipeline for immune repertoire data analysis in their cloud environment. The pipeline had to be:

- extremely efficient to enable the processing of hundreds of millions of reads
- easily scalable to accommodate the ever-increasing amount and diversity of data
- customized regarding processing options and visualizations

Our approach

To assure reproducible and scalable data analysis, our experts built the pipeline using the Snakemake workflow manager. By implementing approaches beyond those available publicly, the BISC team increased data analysis throughput and so enabled the efficient processing of enormous amounts of data sets. With the customized workflow that generates a report easily understandable by the customer, the pipeline is supporting the discovery of novel antibodies via biopanning, i.e., by analyzing the enrichment of specific clones across multiple rounds of binding.

Result and impact

With the ability to discover more novel antibodies faster, our client gained a competitive advantage in the field of antibody therapeutics discovery and development.