Datasets

Case study: Research & Development

Using knowledge graphs to drive epigenetic target discovery in oncology with AstraZeneca

ELSEVIER

Advancing human progress together

Introduction

Learn how by working with Elsevier an Oncology team at AstraZeneca was able to assemble the literature base and build a powerful suite of computational tools to increase the efficiency and depth of target discovery in the epigenetic space, including an order of magnitude increase in the search space for new epigenetic targets.

Challenge

Understanding epigenetic mechanisms in cancer represents a very promising avenue for designing next-generation cancer therapeutics. An oncology team at AstraZeneca wanted to predict novel drug targets in the context of epigenetic regulation of specific cancers.

However, it is difficult to navigate and contextualize the huge and growing body of literature on this subject. Identifying meaningful targets and relationships worthy of further exploration is therefore a challenge.

Solution

Based on specific research questions from AstraZeneca, Elsevier mined data on epigenetic relations from peer-reviewed articles on the ScienceDirect platform. They combined this dataset with a subset of Elsevier's Biology Knowledge Graph, creating a custom knowledge graph called the "EpiMap."

This knowledge graph contextualized these epigenetic phenomena in other known biology and disease processes. The team then built analytical and predictive models based on these data.

Insight

Impact for AstraZeneca

- Elsevier's automatic text-mining workflows resulted in an expansion of the search space for new drug target signals by an order of magnitude which would have been infeasible through manual curation.
- Insight into mechanistic paths of known interactions including new insights into testable pathways of drug resistance identified across the literature.
- Confirmation of known results for four initial research questions and identification of interesting new targets to validate.

Charting a path through the literature

A small oncology R&D team at AstraZeneca was researching several cancer subtypes with the aim of identifying new epigenetic drug targets.

Epigenetic changes are prolific in cancer. Unlike genetic changes, epigenetic phenomena can be reversed through pharmacological intervention. This makes them a potentially rich source of new drug targets.

Thousands of new findings on epigenetic changes in cancer are published every year, so it can be difficult to identify and prioritize new targets for drug discovery. Manually searching this volume of literature is impractical.

Literature mining and knowledge graphs

Researchers need support in the form of tools and processes that automate data harvesting. Moreover, they need a way to contextualize these data in known biological pathways, including pathways involved in disease processes.

An effective way to organize findings mined from the literature is in the form of a knowledge graph. In this context, a knowledge graph identifies biological entities (such as proteins) and their relationships (such as the regulation of other proteins or diseases). It allows researchers to visualize complex biological pathways involved in disease and identify gaps in knowledge.

For drug discovery in oncology, knowledge graphs have some key advantages. The graphs can support various types of data, including genetic and clinical data. This allows researchers to depict the many facets of disease mechanisms and explore complex non-linear biological pathways in accessible and explainable ways.

Making the leap to predictions

Using predictive methods, such as machine learning, in combination with knowledge graphs, researchers can infer links between drugs or proteins. Such links may have been previously unknown, revealing pathways and entities ripe for drug targeting.

If a knowledge graph shows, for example, that two different proteins regulate a disease, and that a drug affects one of these proteins, it may be possible to predict the likely relationship between the drug and the other protein.

These artificial intelligence models can surface missing links in the graph by ranking entities in terms of their likelihood to complete a relationship with another entity. Such newly discovered links can suggest possible gene-disease regulatory relationships, for example.

These rankings can serve as a useful signpost on the drug discovery journey. They could illuminate previously under-appreciated or unknown disease mechanisms and inspire a new direction for research.

Collaboration and transparency to meet a challenge

The team at AstraZeneca was interested in knowledge graphs harvested from scientific literature as a means to unlock new drug targets, but they did not have the capacity to fully pursue this method of exploration.

Through conversations with Elsevier AstraZeneca concluded that Elsevier had the scientific domain knowledge, data, and data science skills to help them pursue this method. Realizing this, they began a datathon collaboration with Elsevier, to develop a knowledge graph and supporting computational ecosystem. They hoped this would allow exploration of key questions in their research area, including identifying new epigenetic drug targets.

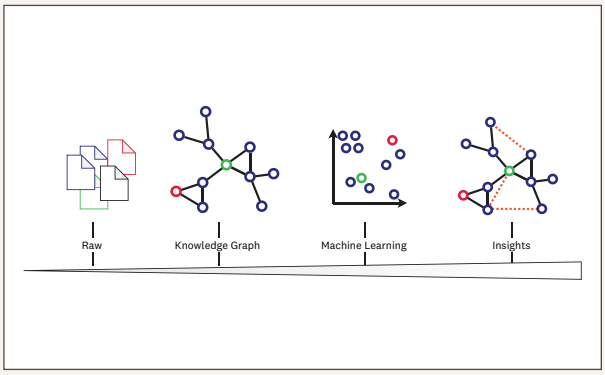
A major challenge in building predictive models in drug discovery is the ability to support novel predictions (unknown unknowns) with an explainability framework. To build confidence in the method, known unknowns must be reliably reproduced before novel predictions can be confidently considered.

Solution

Creating and populating EpiMap

The Elsevier team began by extracting relevant data from peer-reviewed articles on the ScienceDirect platform and organizing the information using natural language processing (NLP). NLP allows computers to read and derive meaning from human-created text. Elsevier combined this bespoke data set with a subset of its Biology Knowledge Graph to create a custom knowledge graph. The resulting EpiMap identified the biological components and relationships that describe how certain cancers arise and develop.

Using this graph as a foundation, AstraZeneca and Elsevier collaborated during a three-month datathon period, addressing a set of challenging and largely open questions in the field of epigenetics in oncology.



Conceptual overview of the process of knowledge transformation in EpiMap datathon.

New tools and training

In preparation for the datathon event, Elsevier spent six months constructing an AI workbench that combined computational resources, data visualization tools and the EpiMap data itself. They trained AstraZeneca's team to use these resources to conduct advanced analysis of the EpiMap. AstraZeneca's researchers found that the tools fit their workflow and were accessible, despite their varied backgrounds in data science.

Research questions

The following four research questions concerning epigenetic changes in specific cancers informed the data search and predictive models (details are confidential).

1. What are the genes that are hypo/hyper-methylate across indications?
2. What are the genes co-sensitive to antineoplastics?
3. What are the potential novel drivers of antineoplastic drug resistance?
4. What novel epigenetic targets can be predicted for seven specific cancer indications?

The researchers addressed these questions using triage queries, NLP algorithms, and statistical analysis tools. One example would be the defined task to surface drug resistance episodes in relation to antineoplastic drugs. To address this, Elsevier created an NLP algorithm that would scan the source sentences mentioning relationships between entities, such as drugs or biological molecules, and retrieve those that also mentioned drug resistance.

A feedback loop to increase confidence

The resulting text sources and relationships were reviewed by AstraZeneca's experts who reliably found known drug-resistance mechanisms among the results. In addition, they were excited by the ability of the graph to return multi-hop reasoned pathways of drug resistance.

They collaborated with Elsevier data scientists to probe the flexibility and granularity of the scientific queries the graph could support. This resulted in a highly productive feedback loop, in which AstraZeneca's researchers, as experts in their field, could confirm the validity of datathon findings. This affirmation helped to build the researchers' confidence in the approach.

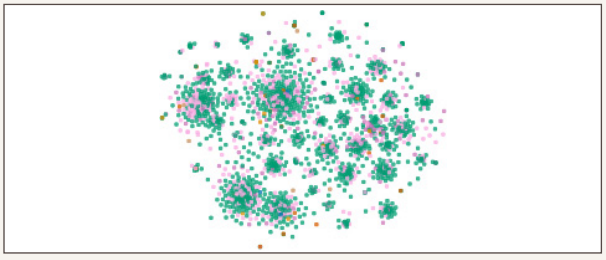
Predicting novel targets

To predict novel drug targets for cancers of interest, the data science team then built machine learning models to predict likely new links in the EpiMap graph. These models employed knowledge-graph-embedding techniques.

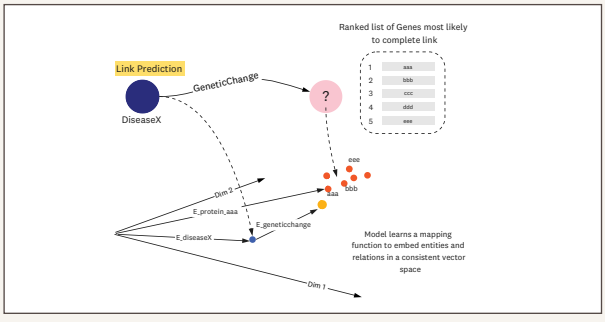
These use mathematical modelling to map the ways in which entities in a graph are related to each other, based on vector representations of entities and relationships themselves. Over three months, the team built and tested such models to discover the most suitable approach to reveal new drug targets involved in epigenetic regulation in specific cancers.

To validate the potential of EpiMap, the machine learning predictions were rigorously evaluated by a panel of scientific subject matter experts from AstraZeneca. Elsevier assisted this validation process by providing further analysis and data visualisations that helped with interpreting the machine predictions. This included the subgraph where a prediction was found along with signals and other supporting evidence to facilitate interpretation.

Projection of subset EpiMap dataset



A view of projection of a subset of entity embeddings in the EpiMap embedding models.



A model was trained by scoring and ranking all possible entities by their likelihood of completing a link.

“EpiMap was critical in driving novel hypothesis generation.”

Director, Oncology Data Science, AstraZeneca

Results

Working with Elsevier, the company gained:

A powerful suite of computational tools to increase the efficiency and depth of target discovery in the epigenetic space. The company’s scientists using these tools describe them as very useful and highly complementary to their workflows.

An expansive knowledge graph constructed from valuable but hard-to-mine data from scientific literature that scrutinizes the role of epigenetic changes in disease processes. The EpiMap confirmed known results and answered new research questions posed by the company’s researchers.

A multitude of significant new insights. These included a number of testable mechanisms of drug resistance across the literature that would have been nearly impossible to identify through manual literature searches. An order of magnitude increase in the search space for new epigenetic drug targets. On average, for each cancer segment studied, the top 10 predictions yielded four known positives (increasing confidence in the approach) and three interesting but unknown candidates. This result highlights the significant power of such methods to unlock new avenues of investigation.

Acknowledgements

AstraZeneca
Vishwa Nellore
Sophie Kirschner
Steven Criscione
Joshua Armenia
James Hadfield
Tom Plasterer
Krishna Bulusu

Elsevier
Payal Mitra
Thom Pijnenburg
Tim Miller
Ted Slater

For more information, talk to your Elsevier Sales Representative today, or visit: elsevier.com/solutions/datasets

ELSEVIER