Drug Discovery as a Recommendation Problem: Challenges and Complexities in Biological Decisions

Anna Gogleva
anna.gogleva@astrazeneca.com
R&D IT, AstraZeneca
Cambridge, UK

Erik Jansson
erik.jansson1@astrazeneca.com
R&D IT, AstraZeneca
Mölndal, Sweden

Greet De Baets
greet.debaets1@astrazeneca.com
BioPharmaceuticals R&D, AstraZeneca
Cambridge, UK

Eliseo Papa
eliseo.papa@astrazeneca.com
R&D IT, AstraZeneca
Cambridge, UK

ABSTRACT

Drug discovery is notorious for its low success rates [5]. Despite best research efforts, the majority of drugs fail at early stages of development, even before they enter clinical trials. This phenomenon stems from the inherent complexity of biological systems and our poor understanding of human diseases. To improve that understanding, swaths of data have been generated in recent years. Still, data does not easily translate into knowledge or actionable insights. Here we explore how approaches from the recommendation system domain could help scientists comprehend the ever-growing amount of biomedical facts. The aim of these efforts is to make better drug development decisions, which ultimately result in safe and efficient treatments for patients [3].

Recommendation systems are well established in e-commerce, streaming and social media platforms, however in the biomedical domain their usage is limited to a few recent studies [1, 6–8]. Direct transfer of classic recommendation approaches to the biomedical domain is not trivial. Specifics of the problem space impose numerous challenges for a recommendation system practitioner, to name a few:

- an elementary unit of recommendation is not a simple self-contained item (like a song or a product), but rather a research direction accompanied by a biologically sound hypothesis;
- ultimate validation of recommendations is complex and often requires expensive and time-consuming laboratory experiments;
- unlike traditional applications, in a biomedical setting both implicit and explicit feedback is at best scarce, making it harder to tune and train models;
- ground truths are rare, and often only relevant for a narrow spectrum of conditions, which renders training challenging;
- context matters, in a sense that each diseases is a heterogeneous and complex molecular phenomenon; depending on what aspect of a disease we want to focus on, different recommendations are expected;
- due to the high cost associated with accepting a recommendation, an increased emphasis is placed on explainability and exposing causal reasoning paths behind a recommendation;
- personalized recommendations approaches are not applicable: in biology or drug discovery decisions are rarely taken by a single user, but rather a team of experts each with different bias;
- previous literature significantly biases users decisions. Users tend to favour known entities, rather than opting for completely novel recommendations (e.g genes residing in the 'dark matter' of the human genome). Adding to this inherent bias, the design of experimental validations is more challenging where less prior knowledge is available.

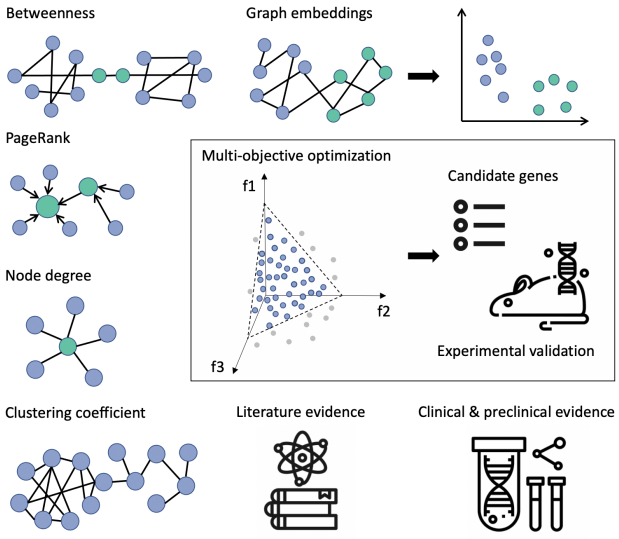

Figure 1: Recommendation system takes into account diverse types of evidence to suggest promising genes driving drug resistance in lung cancer patients. Recommended genes are experimentally validated in disease models.

RecSys '21, September 27-October 1, 2021, Amsterdam, Netherlands

ACM ISBN 978-1-4503-8458-2/21/09.
https://doi.org/10.1145/3460231.3474598

Regardless of the challenges, the adoption of recommendation systems presents numerous opportunities to support and accelerate drug discovery. Even a slight increase in success rate of drug pipelines will result in a vast number of patients gaining access to safe and effective treatments. Recommendation systems could play a leading role in this process.

Adding context to experimental data is one class of problems that could benefit from recommenders. In this process new data is integrated with prior evidence to produce a new hypothesis. In a typical scenario, thousands of genes need to be ranked by their relevance to a disease given new and existing data. As a case study we focused on finding out why some lung cancer patients develop resistance to treatments. Current protocol to find resistance markers starts with high-throughput genomic screens resulting in an initial list of potential gene candidates, followed by tedious manual curation by several experts to reduce the list to a manageable number for further follow-up.

To find resistance markers faster and to reduce bias we built a hybrid recommendation system on top of a heterogeneous biomedical knowledge graph [2]. In the absence of continuous feedback and training data, we approached recommendations as a multi-objective optimization problem [4]. Genes were ranked by trading off diverse types of evidence that link them to potential mechanisms of resistance in lung cancer. We used a knowledge graph as the primary source of features, so that the relevance of a gene could be expressed via properties of a graph. Our hybrid feature set also included clinical and pre-clinical data as well as metrics of literature support obtained with natural language processing techniques. This hybrid approach helped to identify novel resistance mechanisms that could have been overlooked by experts due to inherent bias or limited integration of data. Most importantly, our method reduced the time required to prioritise resistance markers from months to minutes and became a standard procedure for processing genomic screens.

Another class of problems exists around target identification tasks. The idea here is to find a molecular target, often a gene or a protein, that could be modulated with a drug to treat a disease. As the number of potential targets is large, the search space can be reduced using network propagation on a dedicated subgraph that captures the functional relationship between genes. This approach also requires a set of seed genes, defined based on high confidence associations with diseases. Disease preferences are then propagated through the network resulting in a preference distribution for the complete set of genes which is used to reduce the search space.

In contrast to adding context to experiments, a considerable amount of training data is available to support target identification. For instance, both successful and failed clinical trials can act as a useful source of data for target identification. Such a setting warrants use of supervised recommendation systems. A supervised approach, however introduces another machine learning hurdle — trust. Since supervised models are typically "black boxes", their quality must be ascertained indirectly, for example using train-test split and estimating model's performance on the test set. Such quantitative performance metrics often are of little value to a biological expert looking for relevant gene targets. Instead, experts instinctively assess model quality by checking if a list of recommendations contains a handful of expected genes [9].

To simultaneously use biologists' intuitions as training data, while avoiding an overly optimistic trust in model output, we used an ensemble modeling approach. We partitioned training data among multiple models such that each available training gene was omitted from one model's training data. The model was then permitted to assess this previously unseen gene in constructing its final list of recommendations, while training genes were removed from consideration. Each model therefore produced a list of recommendations based on an incomplete set of genes. A final set of recommendations was then constructed by collating each individual model's output list. Because these output lists were constructed with biologist input through supervised training, biologists placed a higher degree of trust in the recommendations. This allowed roughly two dozen genes to be fast-tracked for manual assessment and experimental screening.

In summary, accumulation of large amounts of biomedical data coupled with a need to comprehend and reason about it makes drug discovery an attractive field to apply recommendation techniques. Specifics of the problem space and complexity of biological systems call for efficient recommendation solutions that could operate in unsupervised or weakly supervised settings. At the same time, a strong emphasis on explainability is essential to gain trust of biomedical experts.

CCS CONCEPTS

• **Information systems** → **Recommender systems**; • **Mathematics of computing** → Graph algorithms; • **Computing methodologies** → Optimization algorithms; • **Applied computing** → Systems biology.

KEYWORDS

drug discovery, multi-objective optimization, knowledge graph, natural language processing

ACM Reference Format:
Anna Gogleva, Erik Jansson, Greet De Baets, and Eliseo Papa. 2021. Drug Discovery as a Recommendation Problem: Challenges and Complexities in Biological Decisions. In *Fifteenth ACM Conference on Recommender Systems (RecSys '21), September 27-October 1, 2021, Amsterdam, Netherlands.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3460231.3474598

ACKNOWLEDGMENTS

We thank Gavin Edwards, Michaël Ughetto, David Geleta, Andrej Lamov, Krisha Bulusu, Dimitris Polychronopoulos and Benedek Rozemberczki for technical and scientific support.

SPEAKER BIO

Dr Anna Gogleva is a data scientist in the Biological Insight Knowledge Graph (BIKG) team in AstraZeneca, Cambridge UK. Her current focus is building recommendation systems on top of the internal heterogeneous biomedical knowledge graph to solve biological problems and speed-up drug discovery. Anna's background is in computational biology and bioinformatics, she has worked on a variety of research problems ranging from studying evolutionary patterns of native CRISPR-Cas immunity systems in microbial communities to investigating evolution and transcriptional dynamics of plant-pathogen interactions.

REFERENCES

[1] Clément Frainay, Sandrine Aros, Maxime Chazalviel, Thomas Garcia, Florence Vinson, Nicolas Weiss, Benoit Colsch, Frédéric Sedel, Dominique Thabut, Christophe Junot, et al. 2019. MetaboRank: network-based recommendation system to interpret and enrich metabolomics results. *Bioinformatics* 35, 2 (2019), 274–283.

[2] David Geleta, Andriy Nikolov, Gavin Edwards, Sebastian Nilsson, Richard Jackson, Anna Gogleva, Vladimir Poroshin, and Eliseo Papa. 2021. Biological Insights Knowledge Graph. (2021). in preparation.

[3] Anna Gogleva. 2021. Drug Discovery as a Recommendation Problem: talk materials. https://astrazeneca.github.io/recsys21gogleva/

[4] Anna Gogleva, Dimitris Polychronopoulos, Matthias Pfeifer, Vladimir Poroshin, Michaël Ughetto, Ben Siders, Miika Ahdesmäki, Ultan McDermott, Eliseo Papa, and Krishna Bulusu. 2021. Knowledge Graph-based Recommendation Framework Identifies Novel Drivers of Resistance in EGFR mutant Non-small Cell Lung Cancer. (2021). in preparation.

[5] Paul Morgan, Dean G Brown, Simon Lennard, Mark J Anderton, J Carl Barrett, Ulf Eriksson, Mark Fidock, Bengt Hamren, Anthony Johnson, Ruth E March, et al. 2018. Impact of a five-dimensional framework on R&D productivity at AstraZeneca. *Nature reviews Drug discovery* 17, 3 (2018), 167–181.

[6] Makbule Guclin Ozsoy, Tansel Özyer, Faruk Polat, and Reda Alhajj. 2018. Realizing drug repositioning by adapting a recommendation system to handle the process. *BMC bioinformatics* 19, 1 (2018), 136.

[7] Tijana Radivojević, Zak Costello, Kenneth Workman, and Hector Garcia Martin. 2020. A machine learning Automated Recommendation Tool for synthetic biology. *Nature communications* 11, 1 (2020), 1–14.

[8] Chayaporn Suphavilai, Denis Bertrand, and Niranjan Nagarajan. 2018. Predicting cancer drug response using a recommender system. *Bioinformatics* 34, 22 (2018), 3907–3914.

[9] Fan Zhang, Ke Zhou, Yunqiu Shao, Cheng Luo, Min Zhang, and Shaoping Ma. 2018. How Well do Offline and Online Evaluation Metrics Measure User Satisfaction in Web Image Search?. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. ACM SIGIR, Ann Arbor MI USA, 615–624.