

Drug Discovery as a Recommendation Problem

Anna Gogleva, Erik Jansson,
Greet De Baets, Eliseo Papa

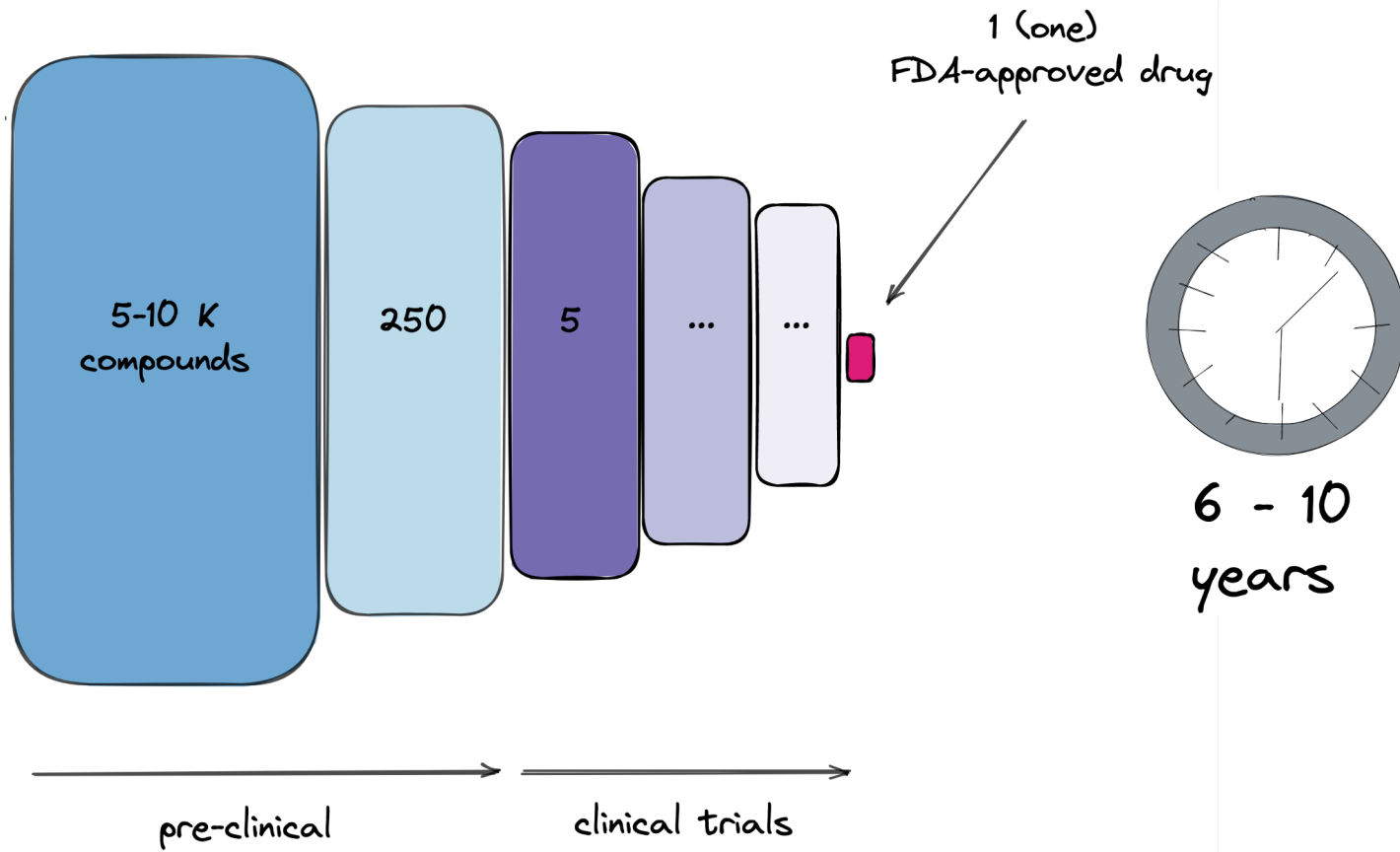
27th September 2021

ACM RecSys'21 Amsterdam

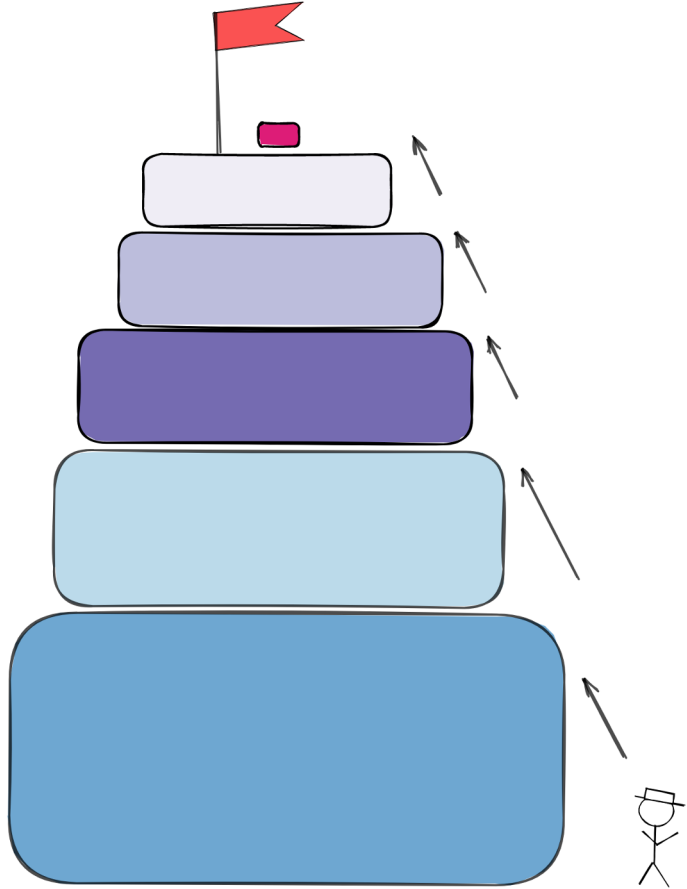
<https://astrazeneca.github.io/recsys21gogleva/>



One needs to fail a lot to discover a working drug



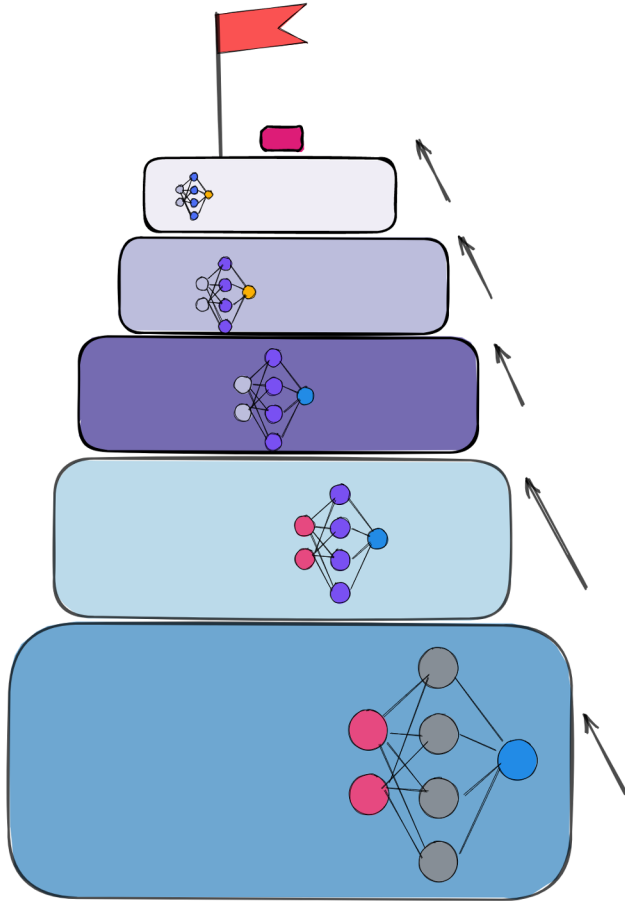
It is a tall mountain to climb



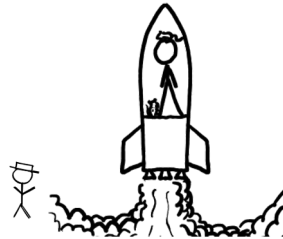
- How to develop new efficient treatments faster?
- How to make better decisions in the process?



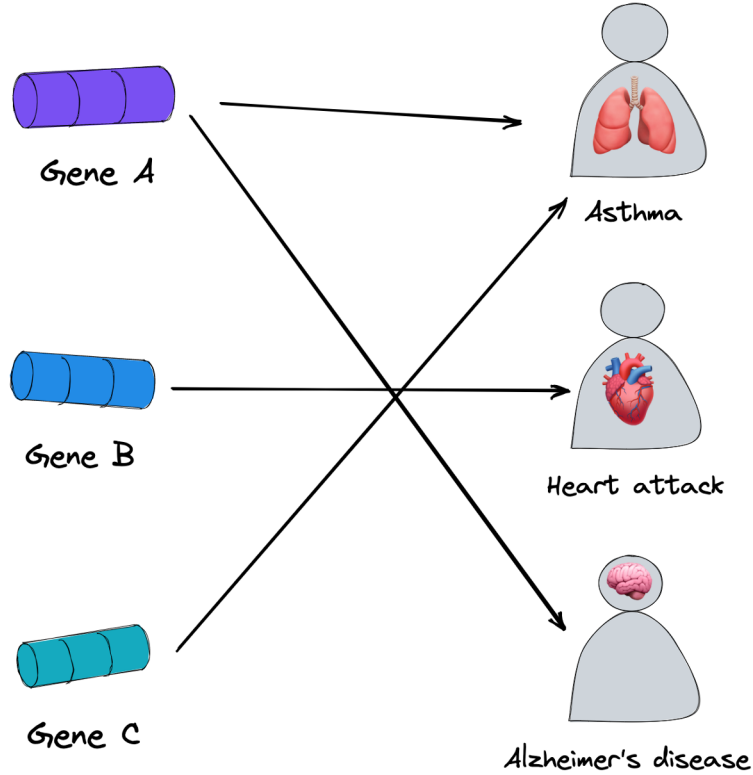
It is a tall mountain to climb



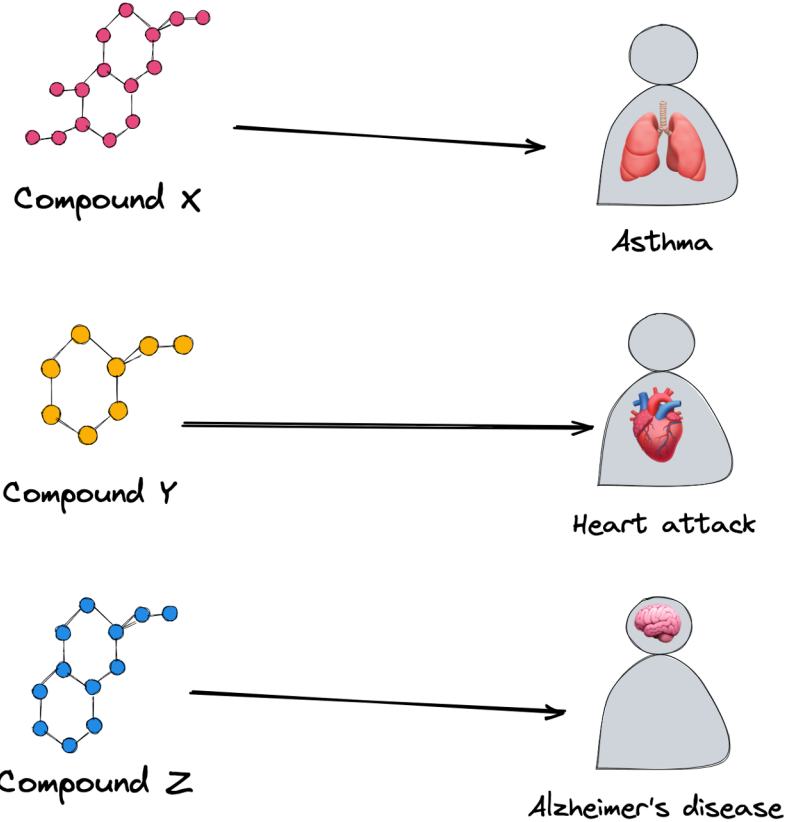
- How to develop new efficient treatments faster?
- How to make better decisions in the process?
- Recommendation systems can help in multiple places



Recommendation problems in drug discovery



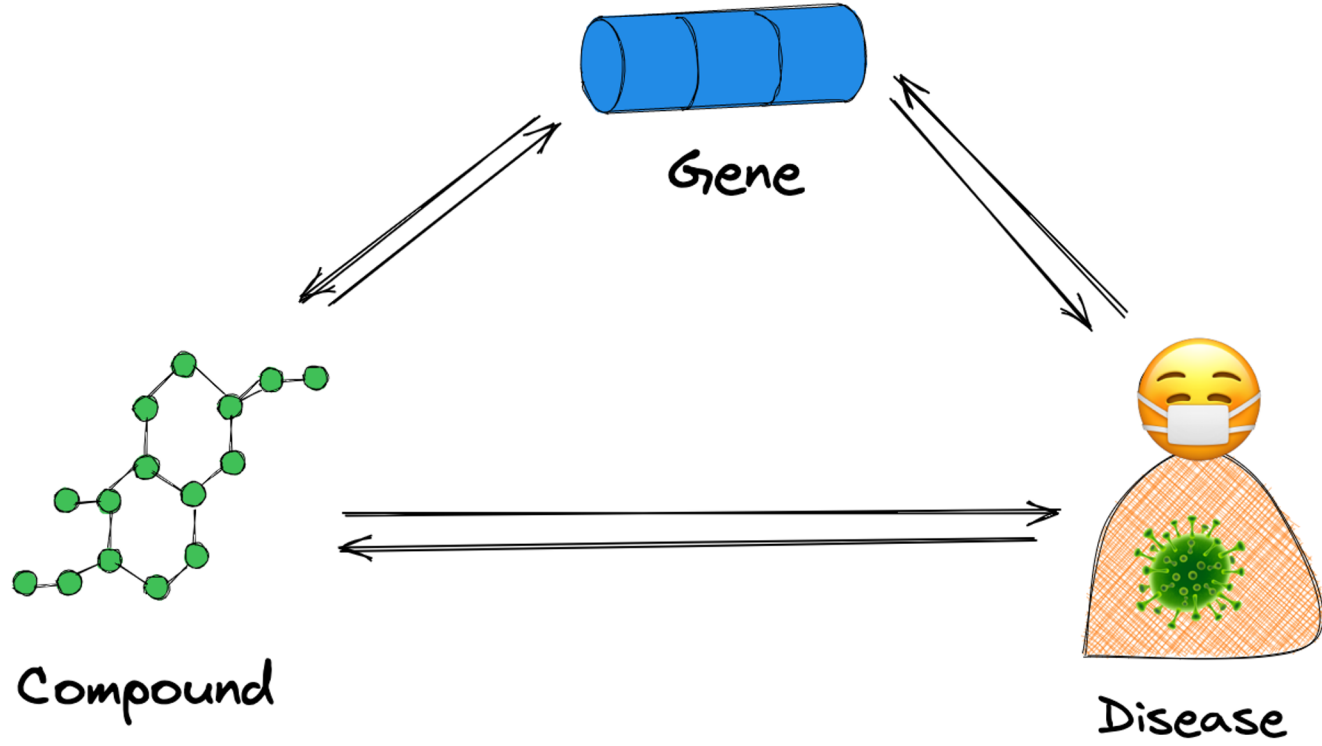
find a gene causing a disease



match a drug with a disease



Drugs, genes, diseases

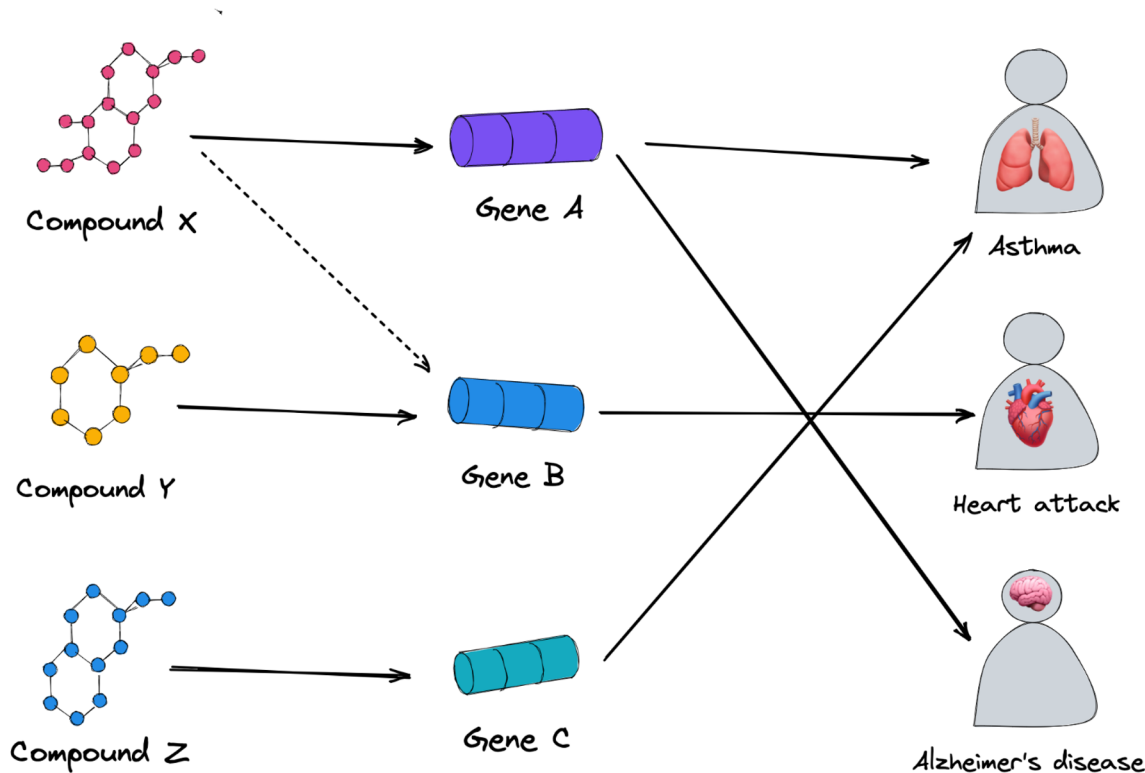


It gets complex very fast

Millions of compounds
Billions possible theoretically

25-30 K genes,
80 K functional elements

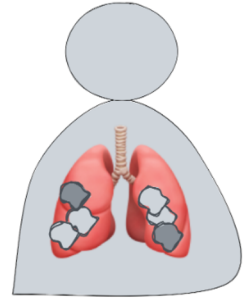
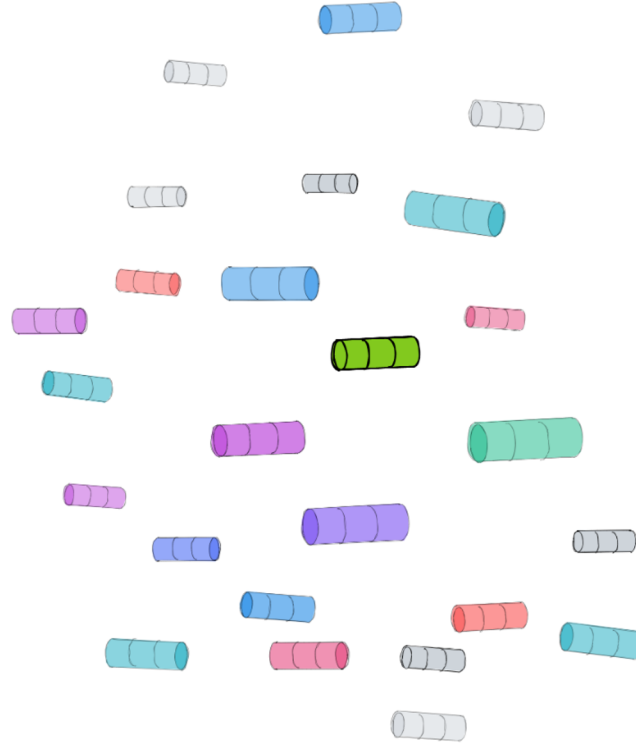
~10 K diseases



It is rarely just a single gene

- 25-30K human genes

- everything interacts with everything,
each gene is a suspect

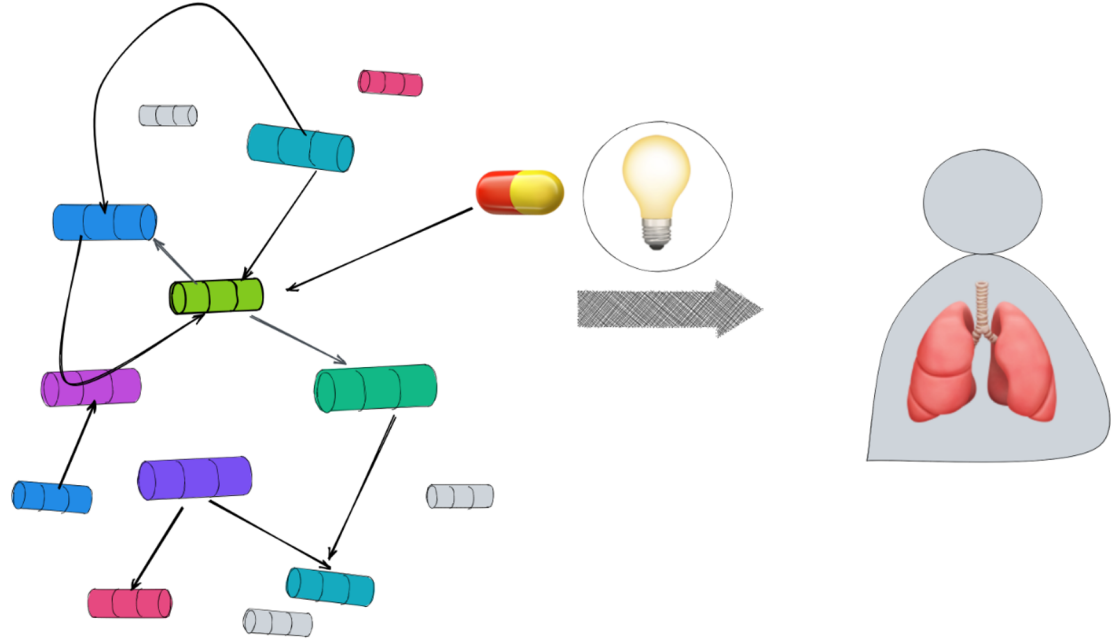


a disease










Find a molecular network behind a disease





- 1 disease \sim a molecular process gone awry
- 2 find the key molecular process
- 3 re-route it safely




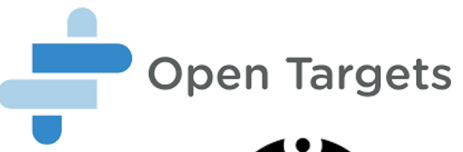



Biomedical knowledge is spread across multiple resources



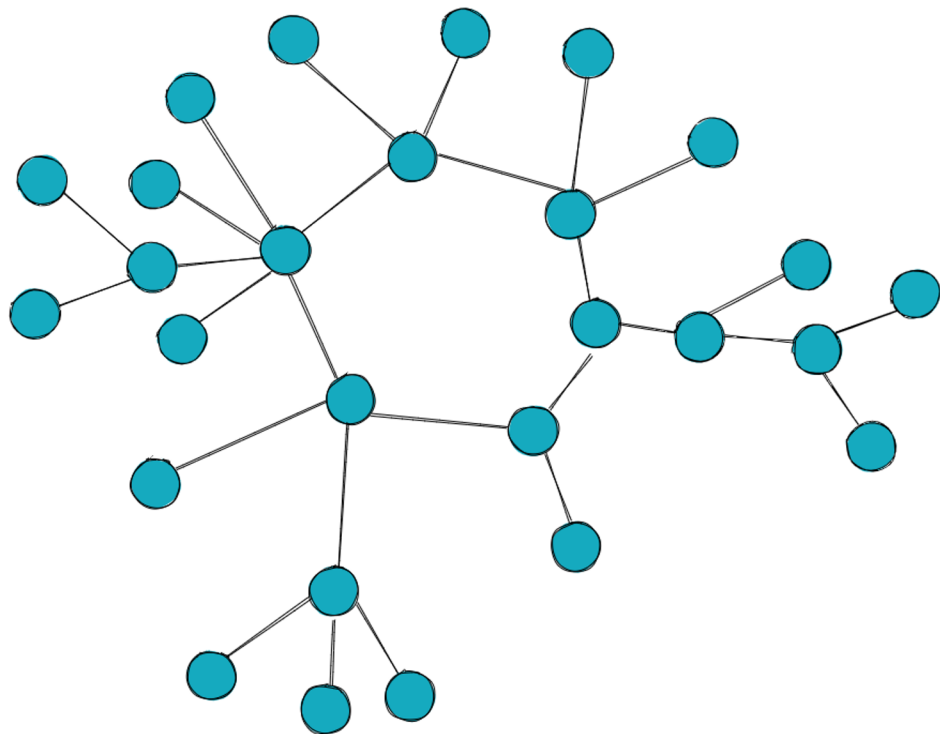








Graph makes things simpler



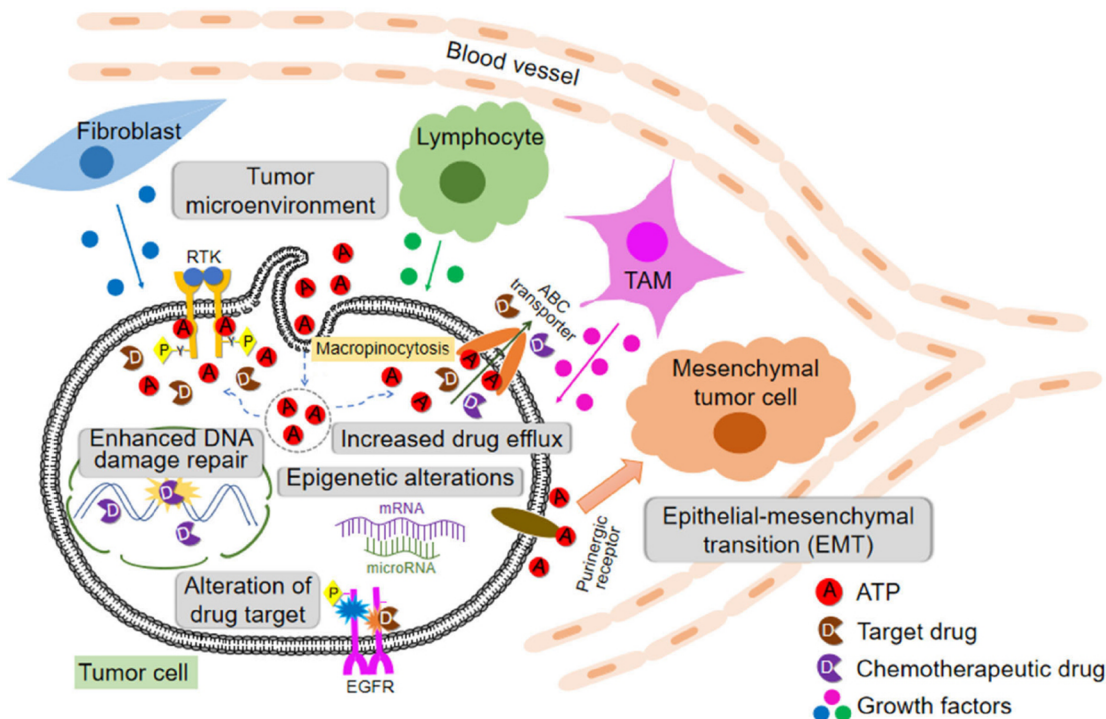
- Biomedical information often comes in forms of **networks** and **hierarchies**
- Graph is a convenient way to organise it
- BIKG (our internal knowledge graph): **60+** data sources including - omics and data extracted from the literature
- **11 M nodes, 1 B edges**
- Use graph as a source of context and features for recommenders



Early success story:

graph-based
recommendations

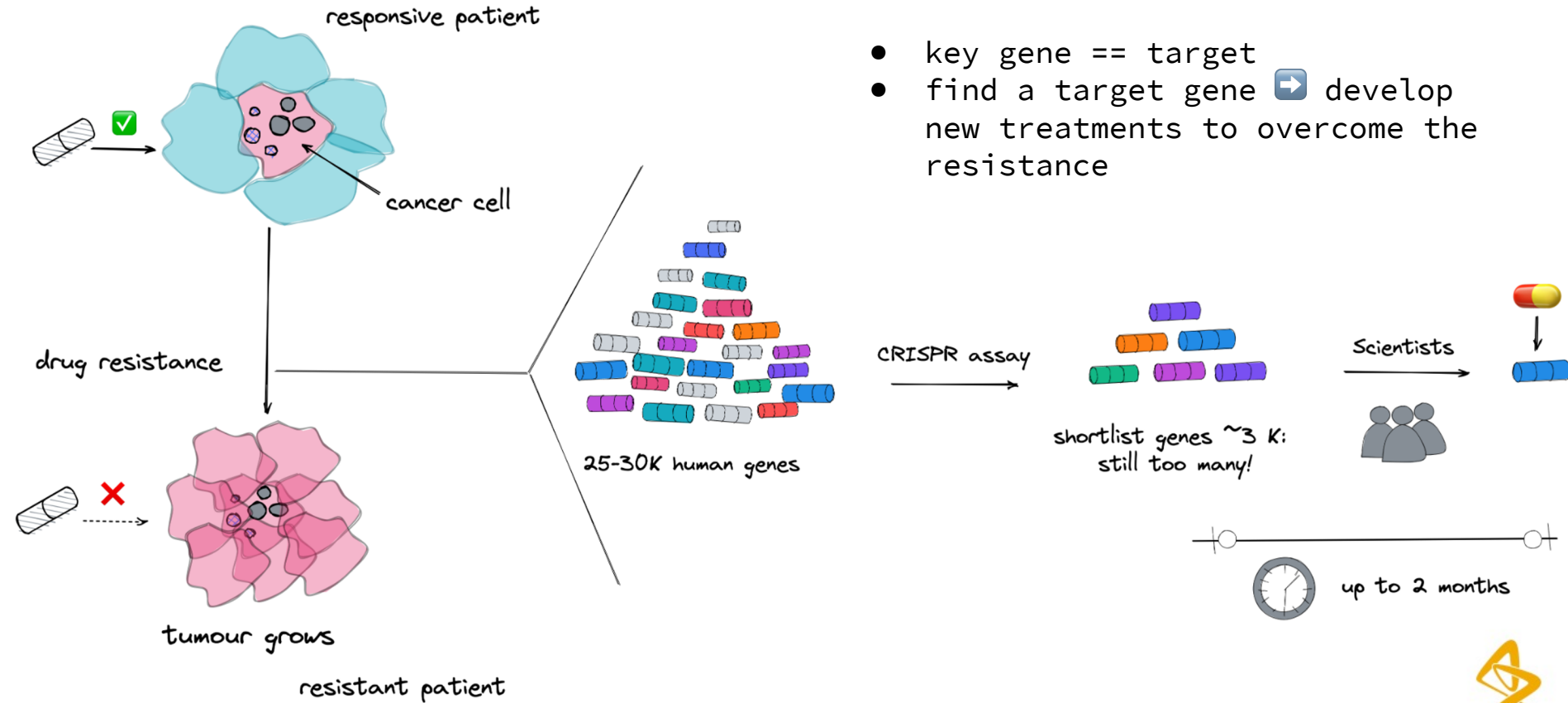
Applied recommendation problem #1: contextualize experimental data



- Drug resistance in lung cancer
- Occurs in a sub-population of patients
- Resistance landscape is complex

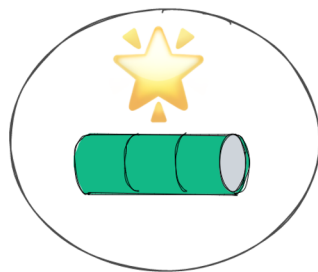


How to help scientist find key genes faster?



An ideal target

— — —



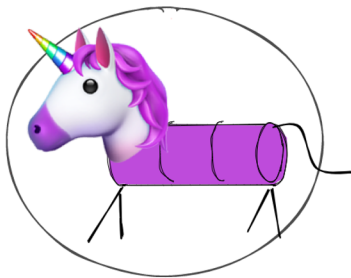
- ☒ Expression
- ☒ Pathway/complex enrichment
- ☒ Effect size
- ☒ Druggability
- ☒ Mode of action
- ☒ Translation in models
- ☒ Internal assets
- ☒ Bench validation
- ☒ Consistency in assays
- ☒ Clinical relevance
- ☒ Literature support
- ☒ Novelty

...



An ideal target does not exist

— — —



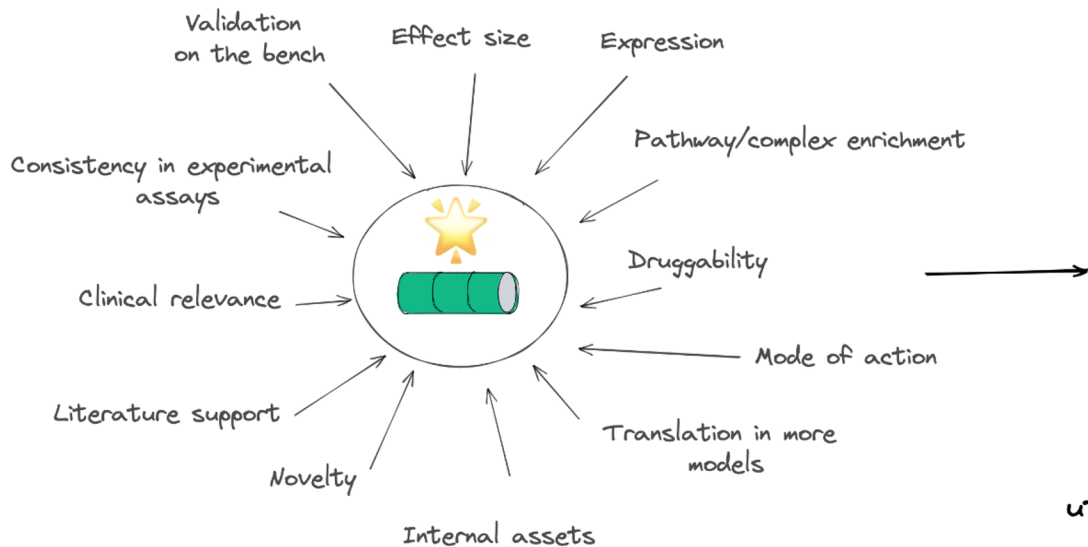
- ☒ Expression
- ☒ Pathway/complex enrichment
- ☒ Effect size
- ☒ Druggability
- ☒ Mode of action
- ☒ Translation in models
- ☒ Internal assets
- ☒ Bench validation
- ☒ Consistency in assays
- ☒ Clinical relevance
- ☒ Literature support
- ☒ Novelty

...

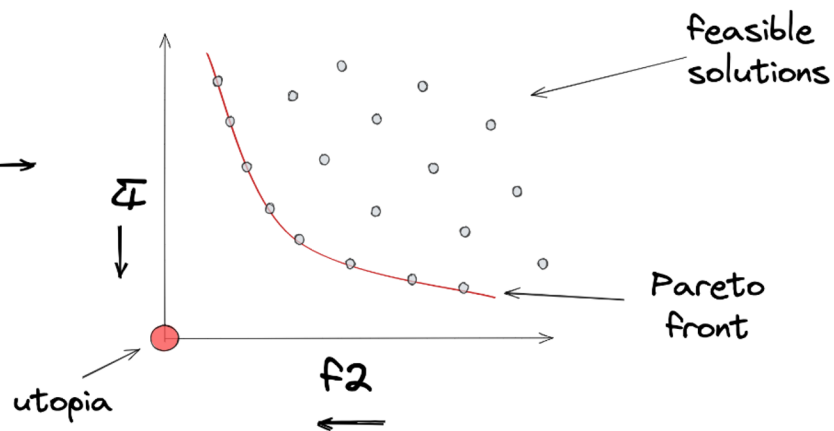


Target selection as an optimization problem

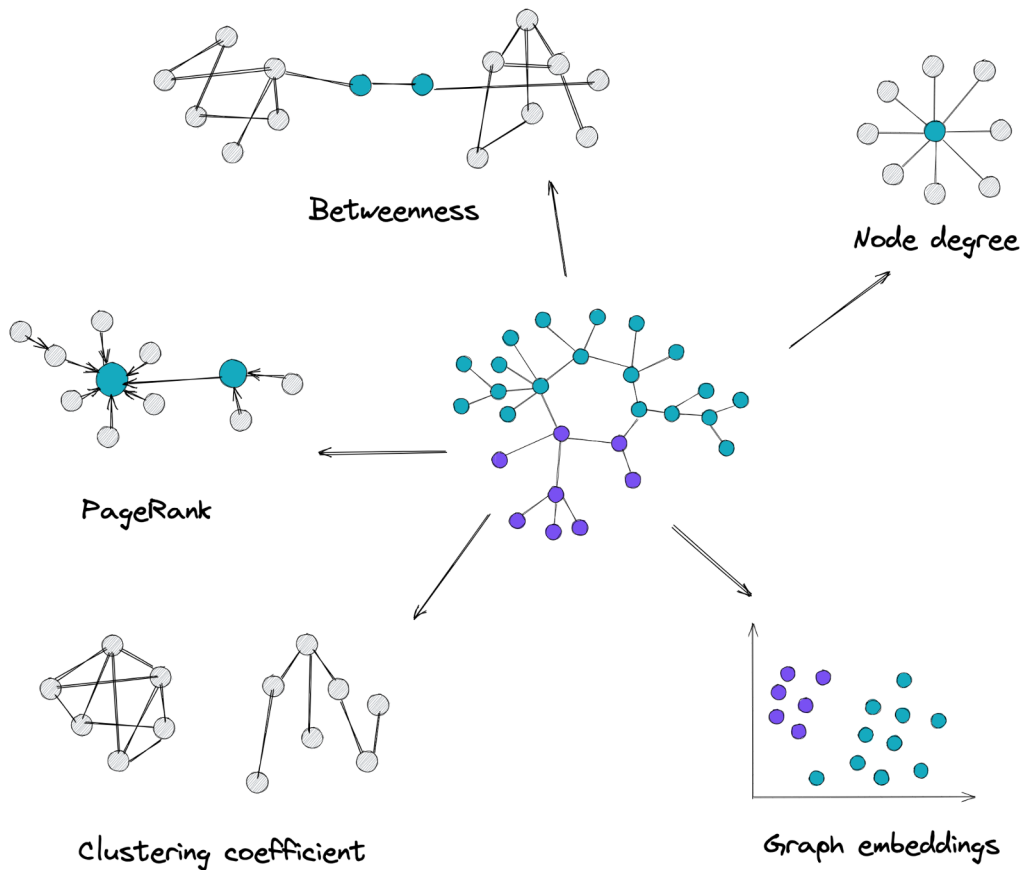
— — —



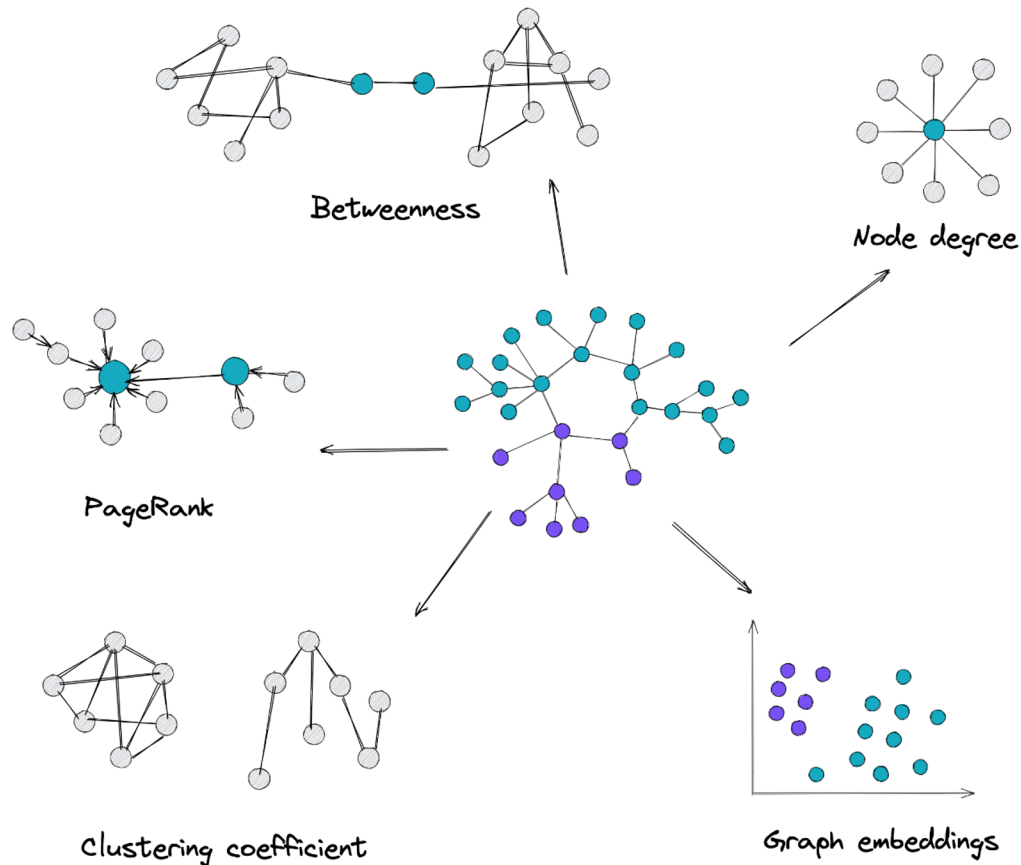
Multi-objective optimization



Hybrid feature set: source features from the graph



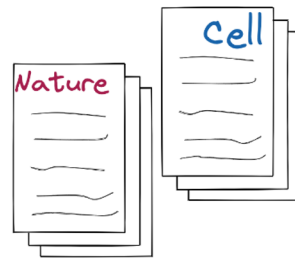
Hybrid feature set: combine with clinical features



clinical features



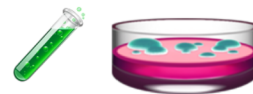
Literature support



Druggability

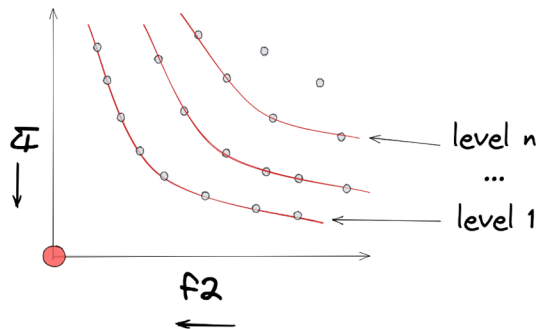


Pre-clinical experimental assays

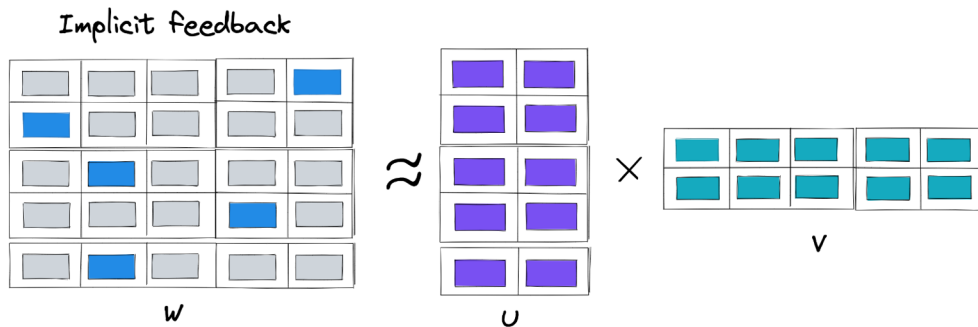


Approaches

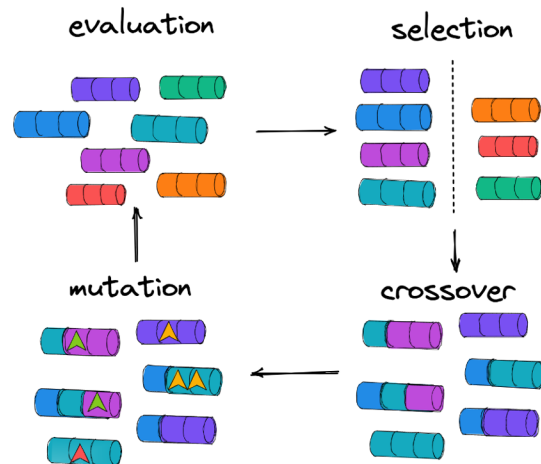
1 Compute exact Pareto front



3 Matrix factorization



2 Evolutionary algorithms



- skywalkR

Essentiality

Tractability

BIKG, literature support

BIKG, graph-derived

Consistency

Clinical relevance

Tesla, enrichment score, RESPONDERS vs RESISTANT

low

high

Flaura, enrichment score, RESPONDERS vs RESISTANT

low

high

Orchard, enrichment score, #ORCHARD vs #FLAURA pre-treatment

low

high

Preclinical evidence

rank!

reset

How it works

Result

Variables explained

Optimal hits, tab view

Sort top genes by

full_screen

Column visibility

Show 10 entries

Search:

gene	trct_sm	nlp_egfr	nlp_nscic	full_screen	KO_osi	KO_gefi	KO_all	A_osi	A_gefi	A_all	tesla_ES
WWTR1	6	0	0	8	0	0	0	5	3	80.00	
KCTD5	0	0	0	8	5	3	8	0	0	0	0.00
NF1	6	0	0	7	4	3	7	0	0	0	0.33
FOSL1	0	0	0	6	0	0	0	4	2	6	0.00
MET	9	8	66	6	0	0	0	4	2	6	0.67
PTEN	6	33	0	6	4	2					
NF2	6	0	0	6	4	2					
CSF1R	9	0	0	5	0	0					
TSC2	0	0	0	5	3	2					
KEAP1	9	7	0	6	2	9					

Showing 1 to 10 of 42 entries

Download Top results

heatmap controls

min cluster size

min number of papers with gene cluster

select a gene, only genes found in NLP clusters are shown

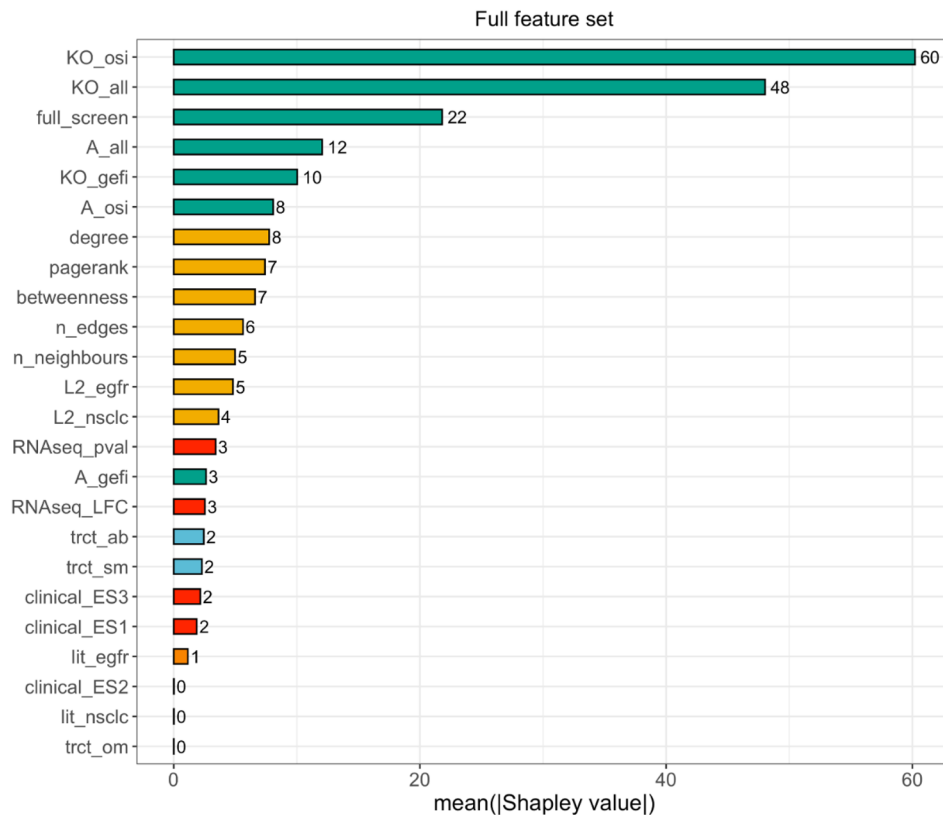
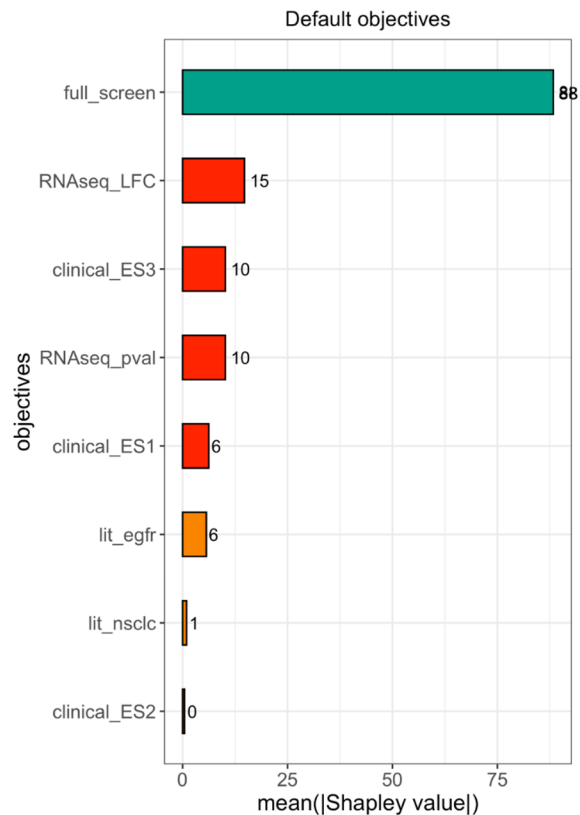
ENSG000001017

heatmap showing multi-term gene co-occurrence in cancer

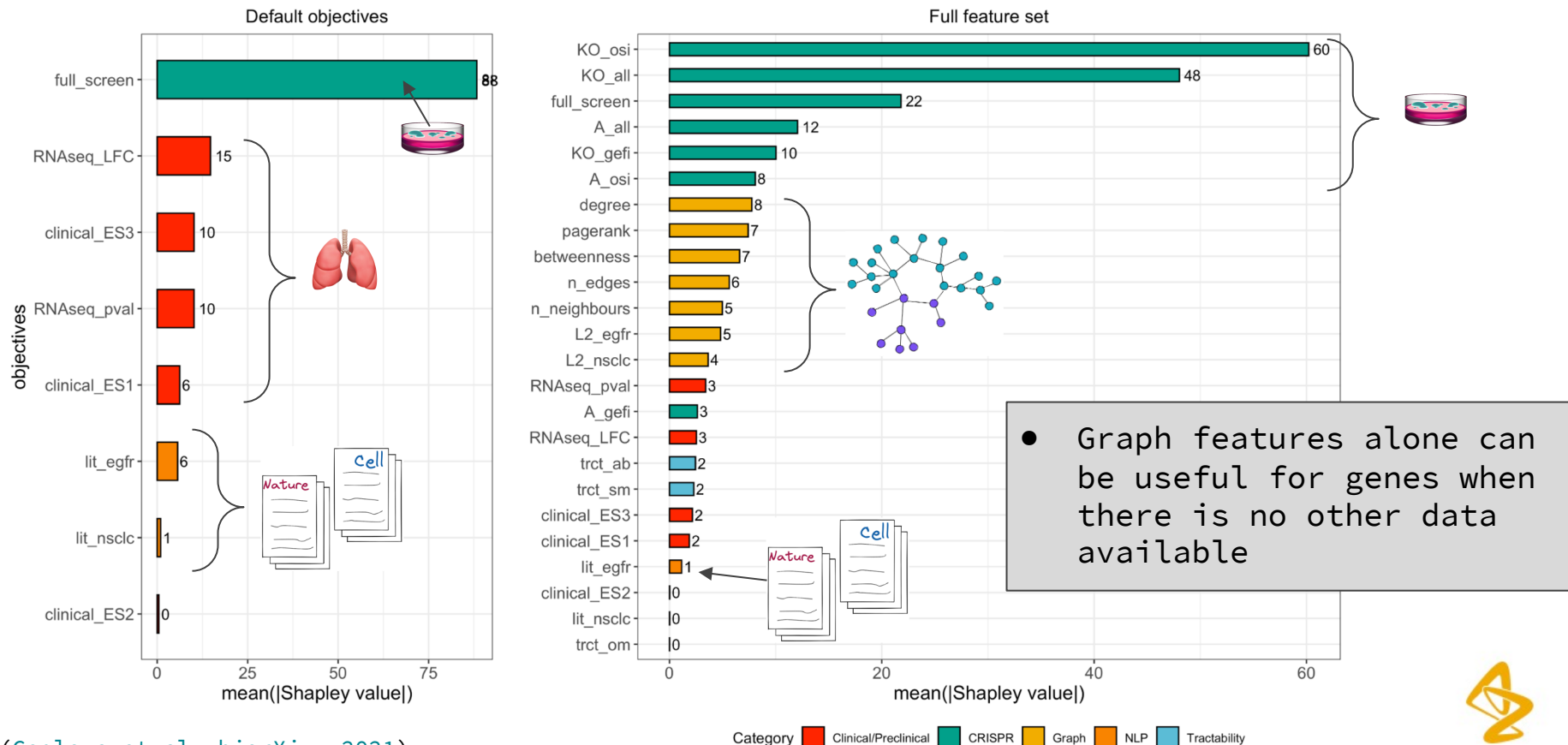


Imperfect validation

Model domain scientist as a black box classifier



Graph-derived features follow clinical in unbiased setting



Annotation by the experts

WWTR1 WW domain containing transcription regulator 1
ENSG00000018408

#Publications of this hit mentioned within the context of 'resistance' and 'EGFR': 0
#Publications of this hit mentioned within the context of 'resistance' and 'NSCLC': 0

for additional evidence behind the gene recommendation please see [skywalk8](#)

☐ Known resistance marker 1

☐ Novel, but credible hit 2

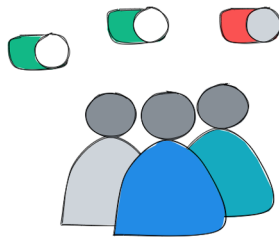
☐ Novel, not credible hit 3

☐ Not novel, not credible hit 4

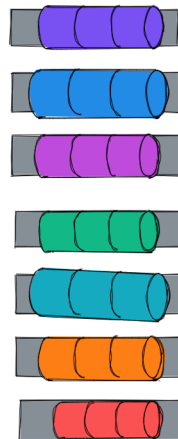
please include any additional details about ongoing experiments for this marker, or if this has been discussed at (pre)TSID.

TASK_NUM: 1 TOTAL_TASKS_NUM: 42

✓ ✗ ⌀ ↶



Gene list



Most of recommendations are 'novel & credible'

WWTR1

WW domain containing transcription regulator 1

ENSG00000018408

#Publications of this hit mentioned within the context of 'resistance' and 'EGFR': 0

#Publications of this hit mentioned within the context of 'resistance' and 'NSCLC': 0

for additional evidence behind the gene recommendation please see [skywalkR](#)

☐ Known resistance marker

1

☐ Novel, but credible hit

2

☐ Novel, not credible hit

3

☐ Not novel, not credible hit

4

please include any additional details about ongoing experiments for this marker, or if this has been discussed at (pre)TSID.

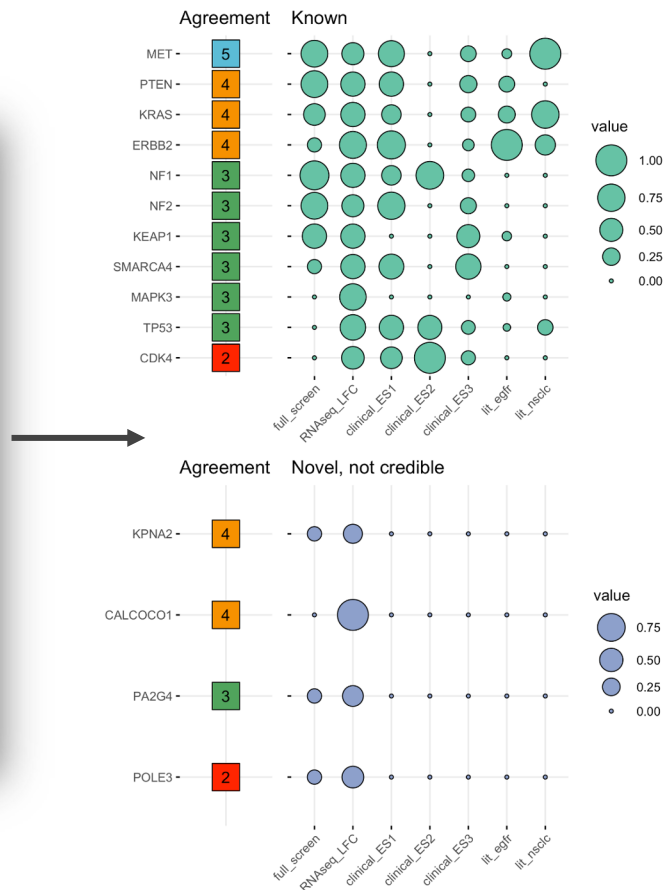
TASK_NUM: 1 TOTAL_TASKS_NUM: 42

✓

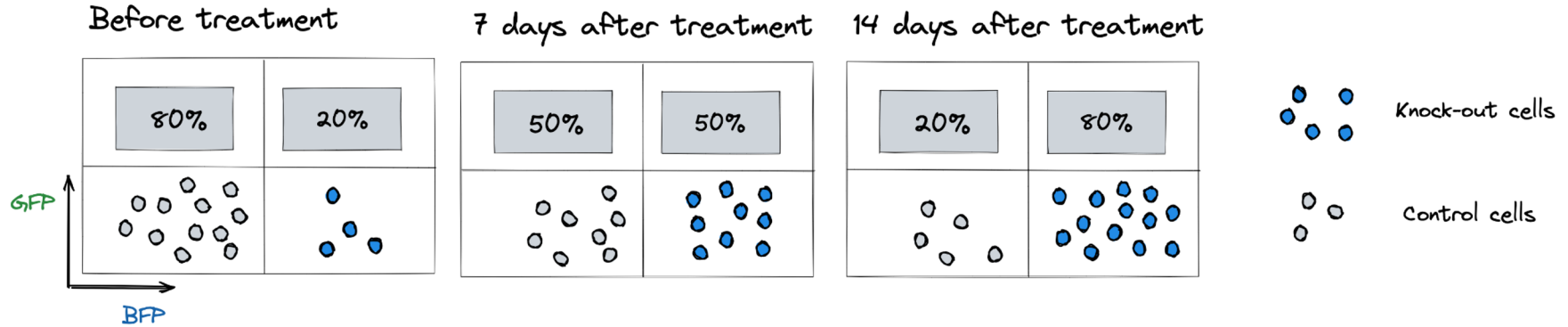
✗

⊘

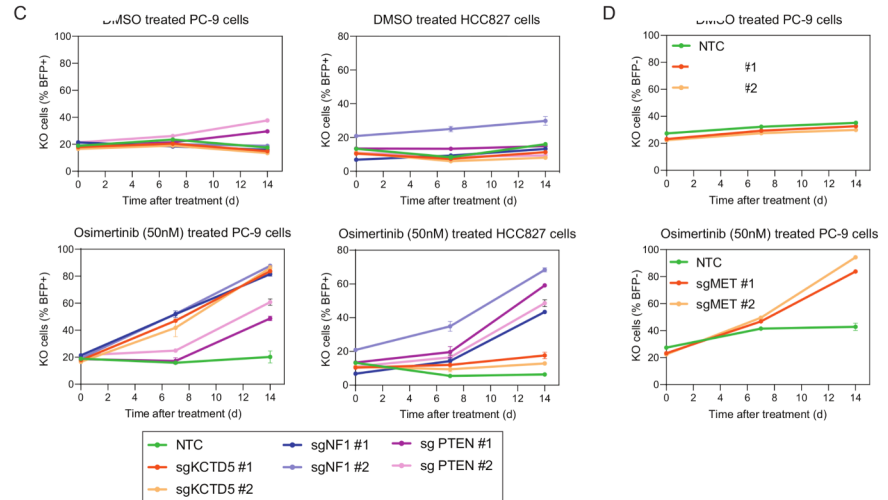
↶



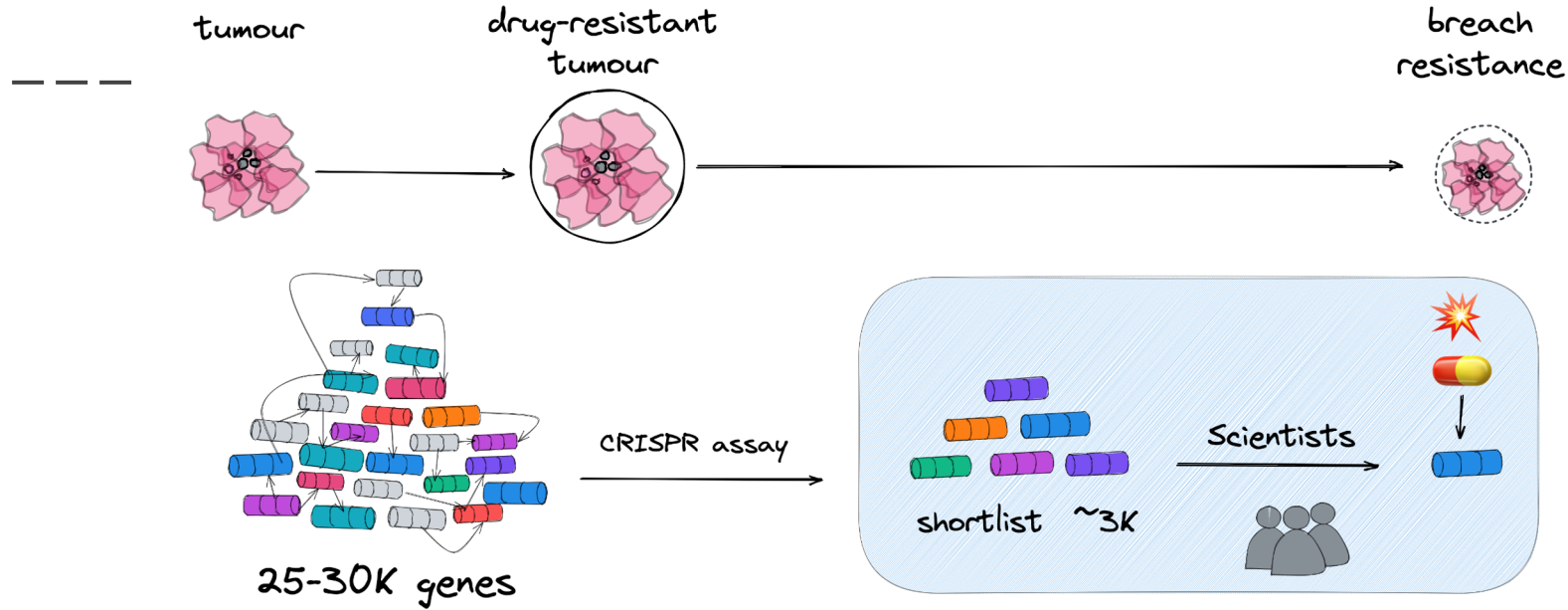
Experimental validation *in vitro*



- confirmed involvement of 4 recommended genes in drug resistance
- next: test the remaining genes



Imperfect, yet already useful recommendation system

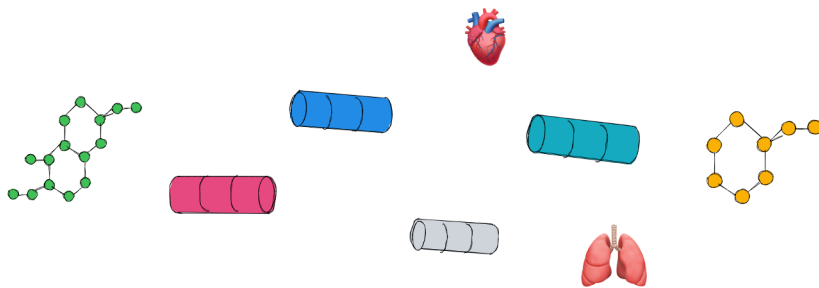


- 🐌 -> 🚗 re-rank lists in seconds, not months
- ⚙️ automated feature generation
- ♻️ approach can be re-used in related problems



Take home message

— — —



- Drug discovery is an exciting field for recommender systems
- Relatively simple recommenders can have a lot of impact
- Need for recommenders that can operate in unsupervised or weakly supervised settings
- There are a number of challenges🟢

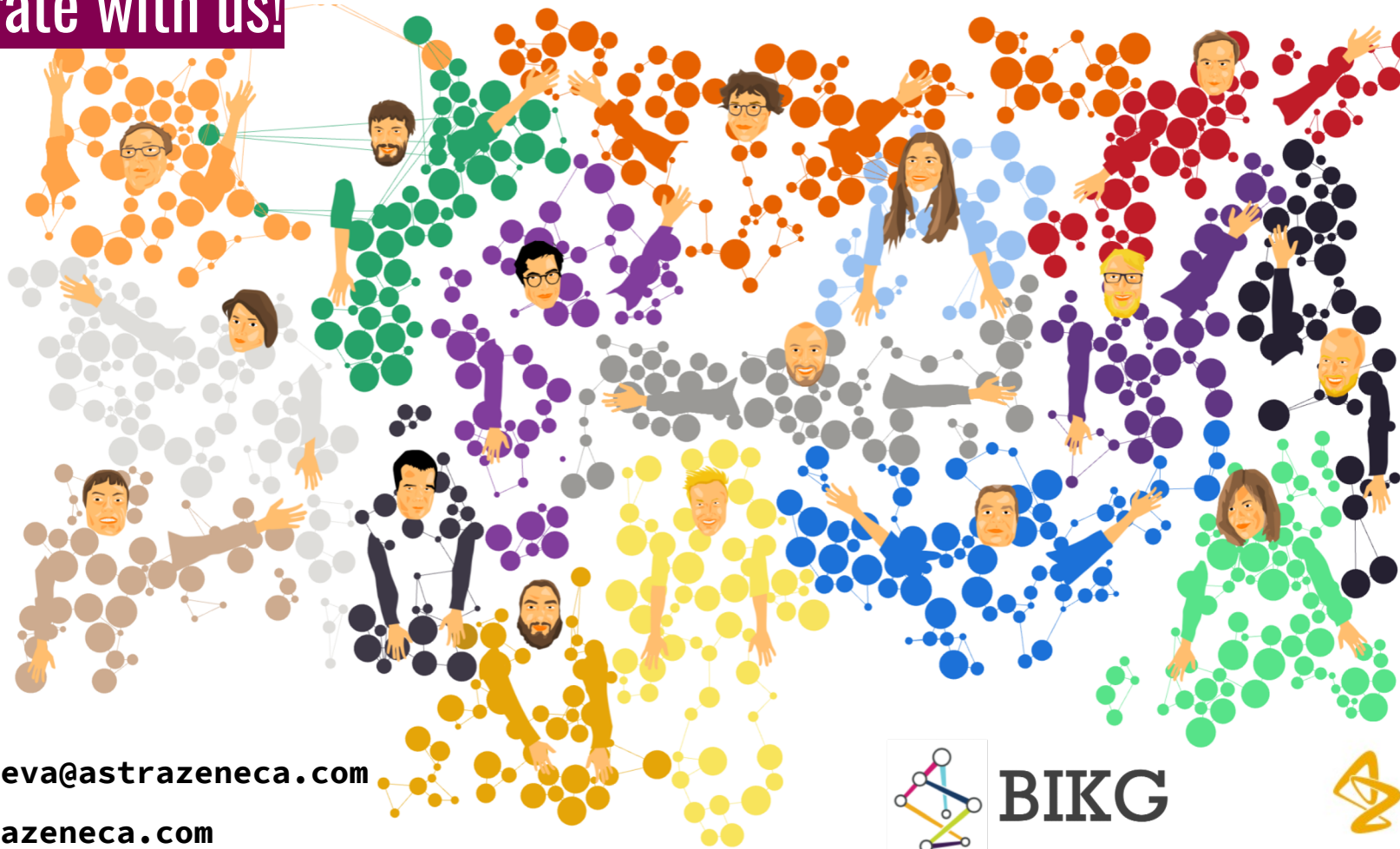


Read more in the extended deck:

<https://astrazeneca.github.io/recsys21googleva/>



Collaborate with us!



anna.gogleva@astrazeneca.com

bikg@astrazeneca.com



BIKG

