# Drug Discovery as a Recommendation Problem

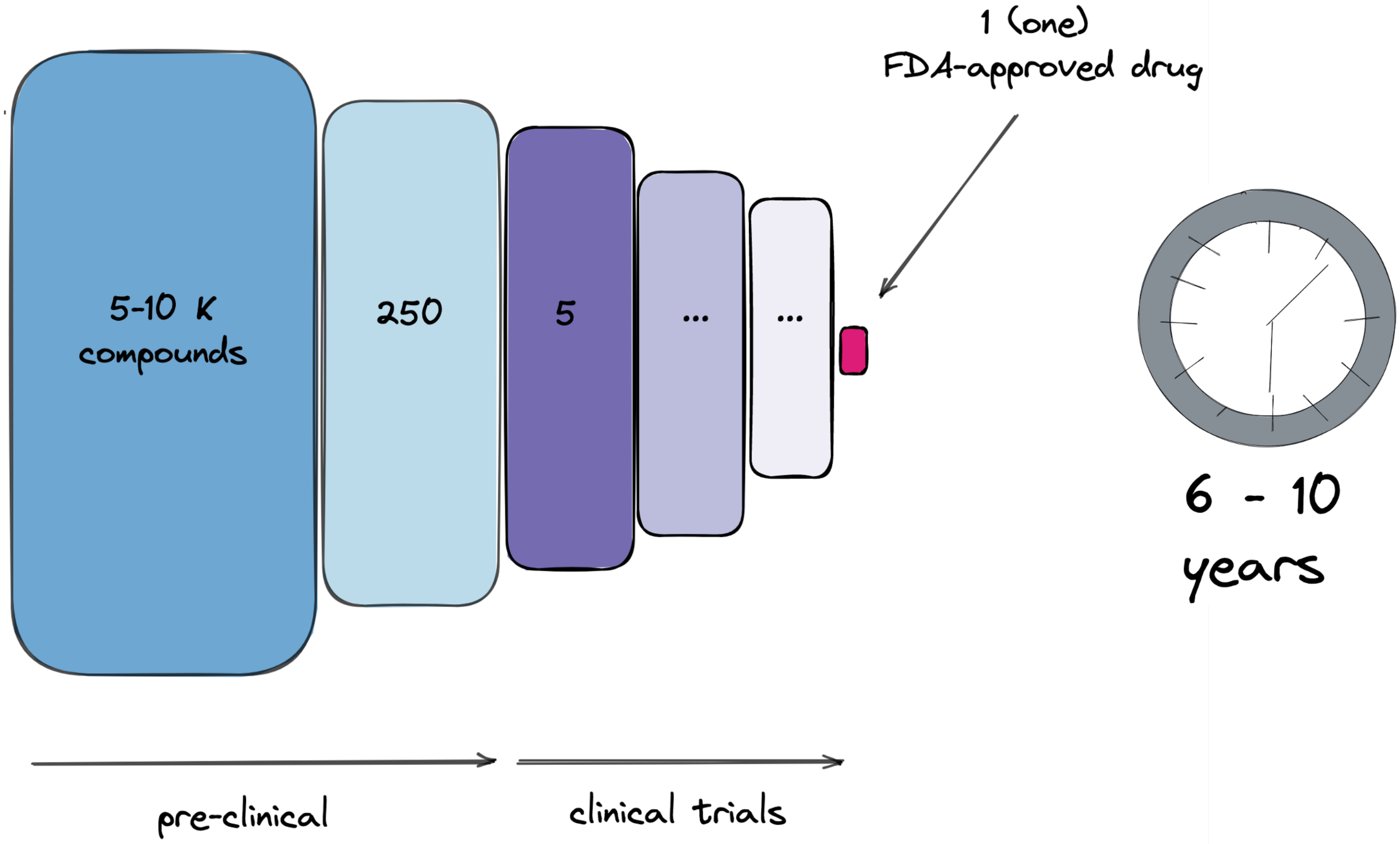
Anna Gogleva, Erik Jansson,
Greet De Baets, Eliseo Papa

27th September 2021

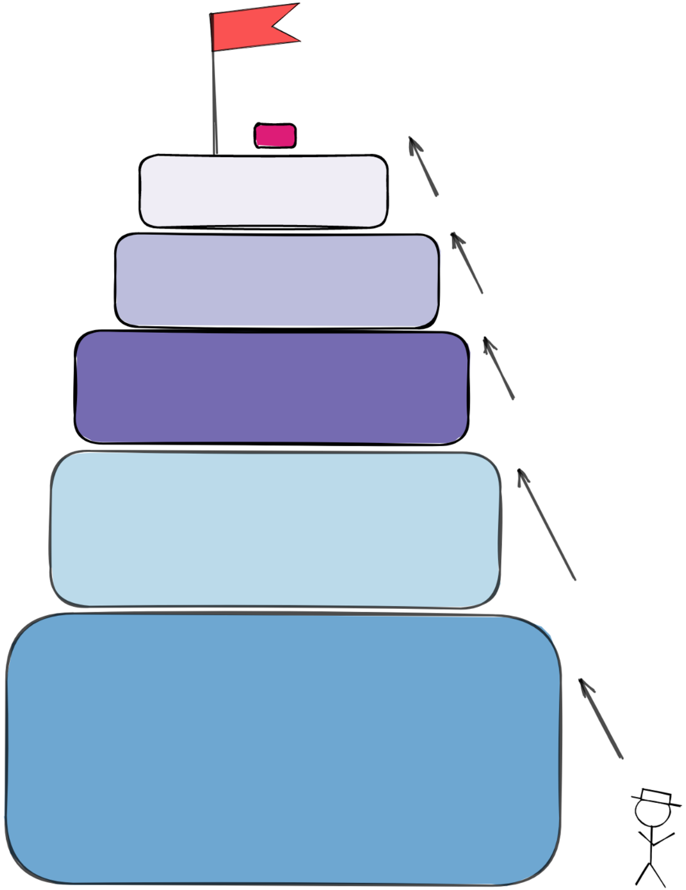
ACM RecSys'21 Amsterdam

**https://astrazeneca.github.io/recsys21gogleva/**

# One needs to fail a lot to discover a working drug

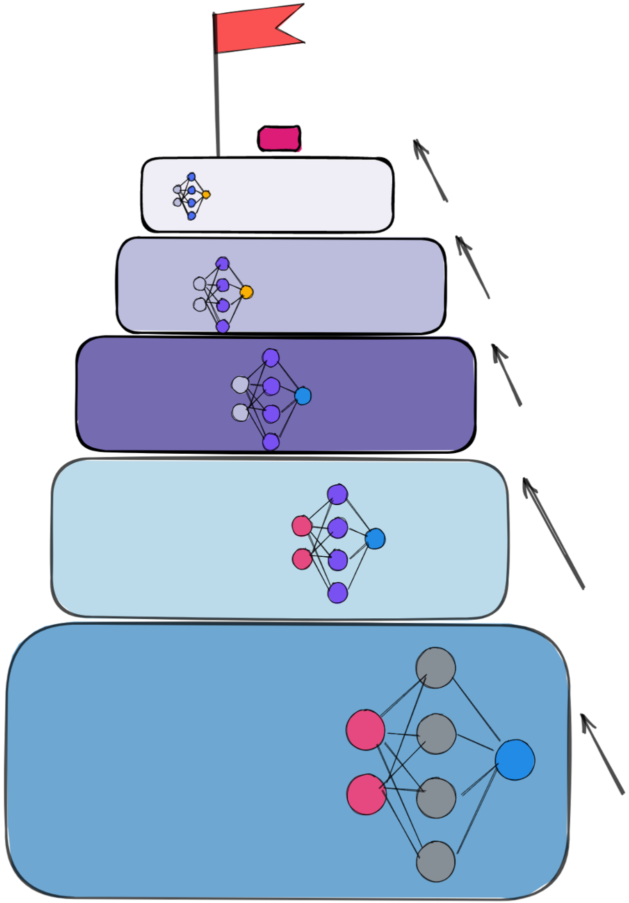


# It is a tall mountain to climb

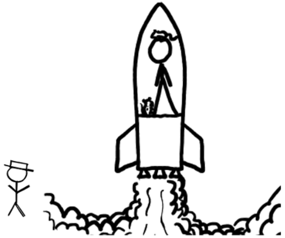


- How to develop new efficient treatments faster?
- How to make better decisions in the process?

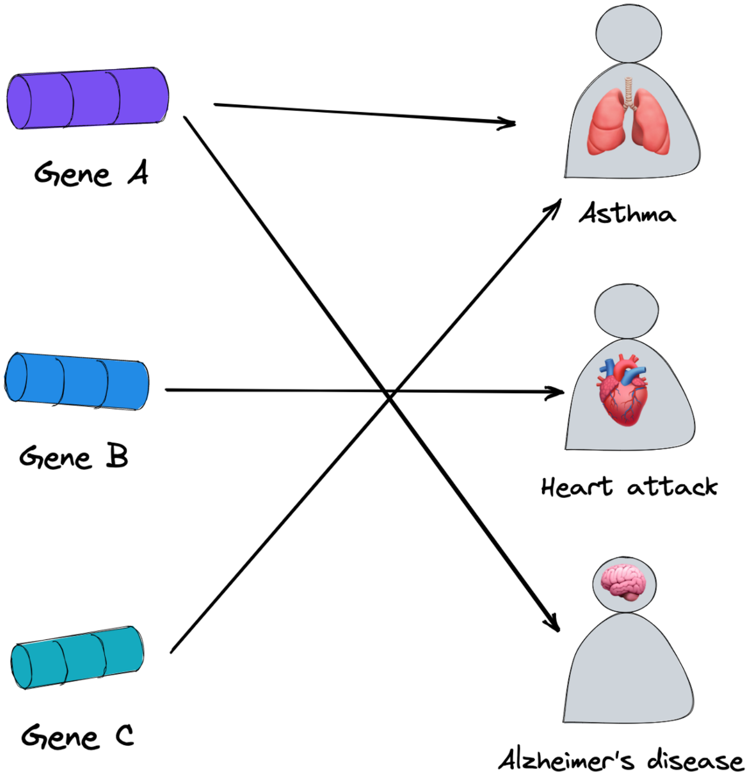
# It is a tall mountain to climb



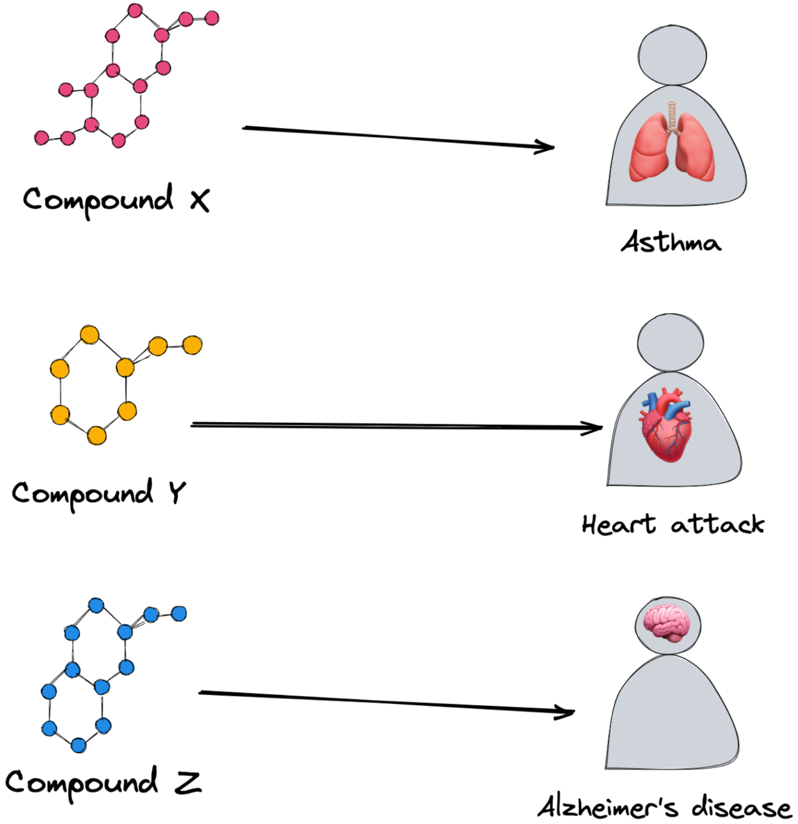
- How to develop new efficient treatments faster?
- How to make better decisions in the process?
- Recommendation systems can help in multiple places



# Recommendation problems in drug discovery

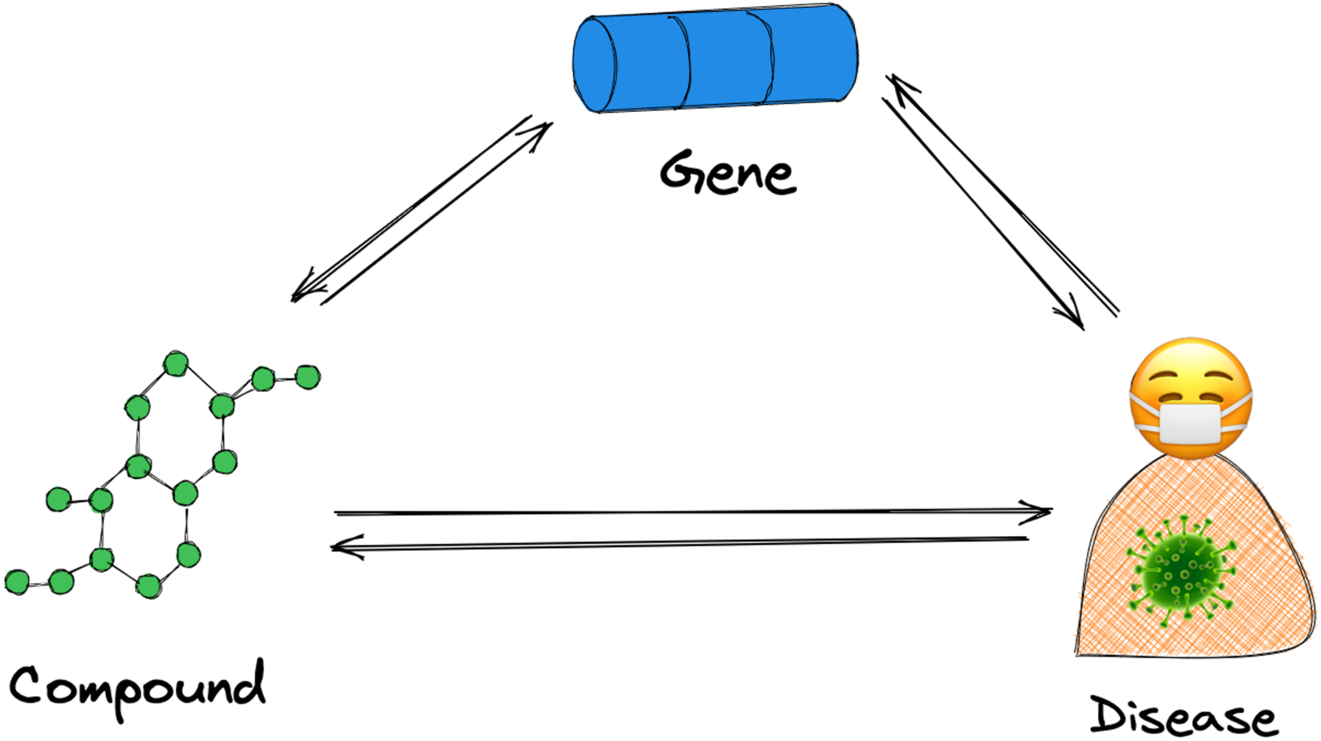


find a gene causing a disease



match a drug with a disease

# Drugs, genes, diseases

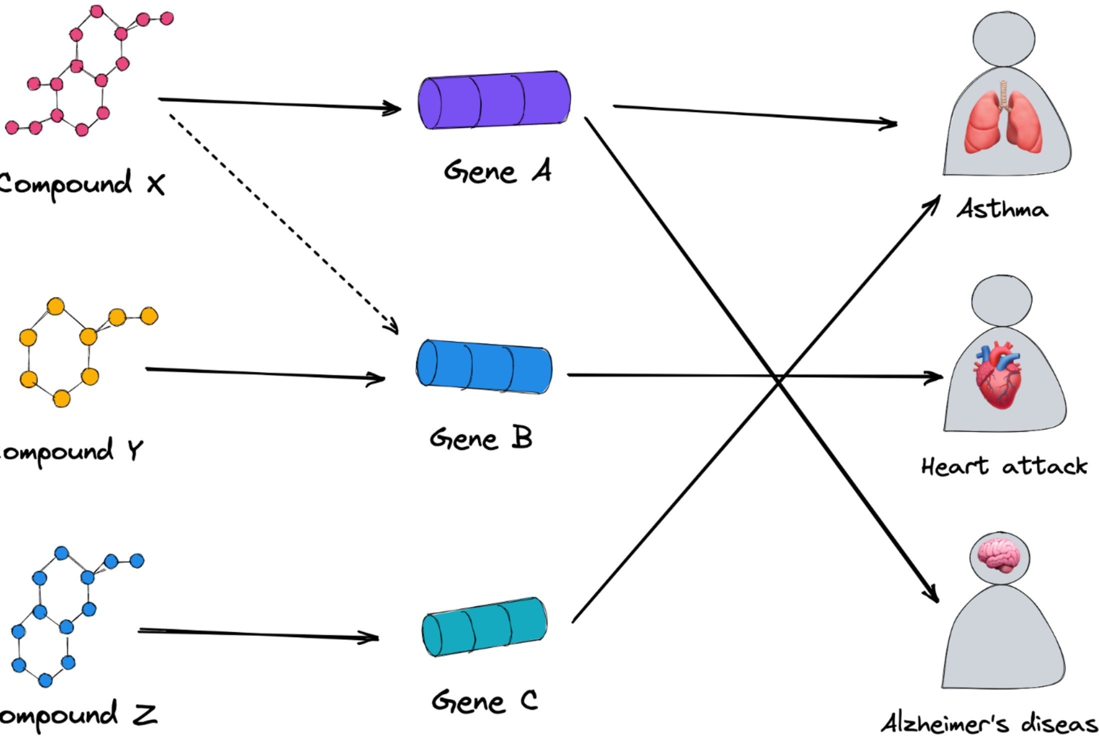


# It gets complex very fast

Millions of compounds
Billions possible theoretically

25-30 K genes,
80 K functional elements

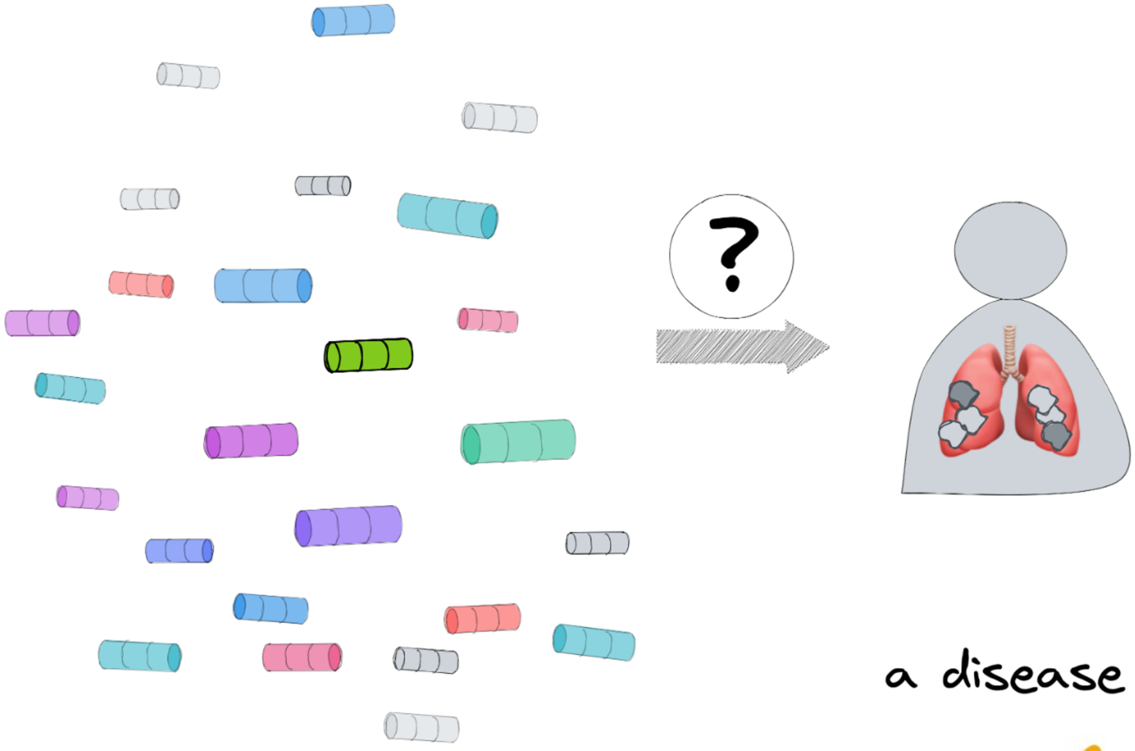
~10 K diseases



# It is rarely just a single gene

- 25-30K human genes

- everything interacts with everything, each gene is a suspect

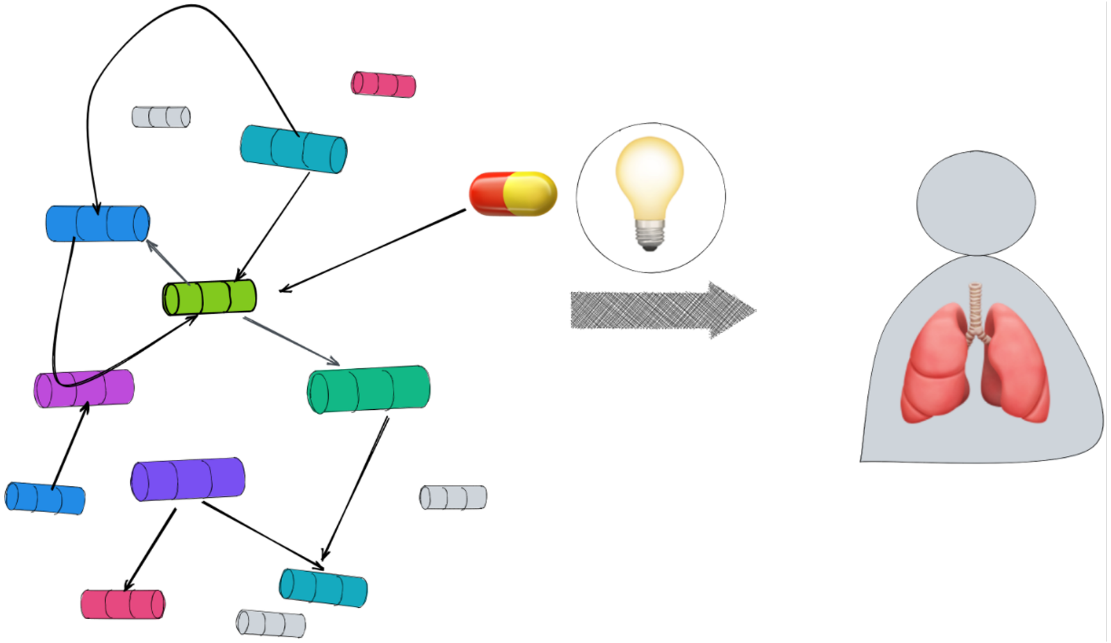


# Find a molecular network behind a disease

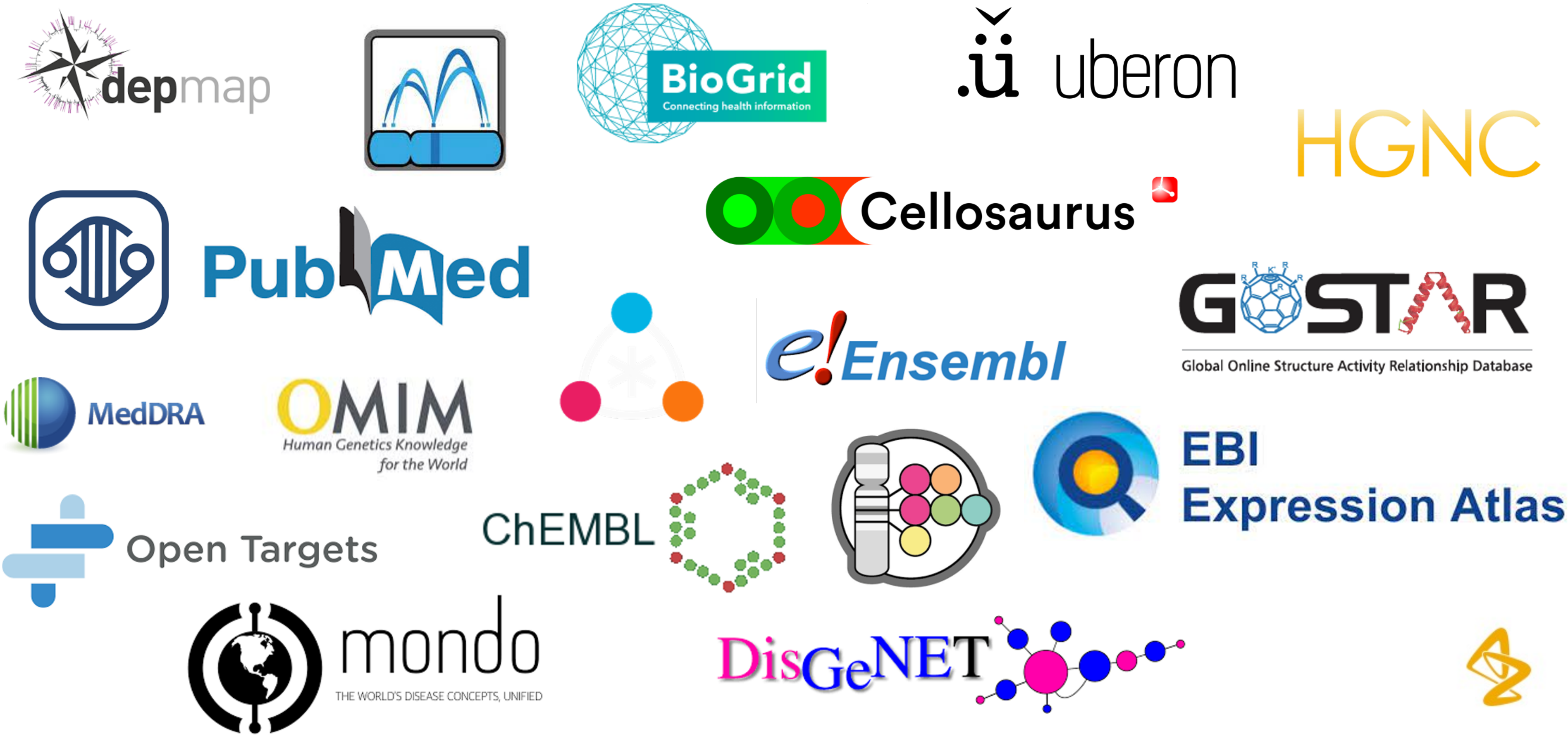
1. disease ~ a molecular process gone awry

2. find the key molecular process

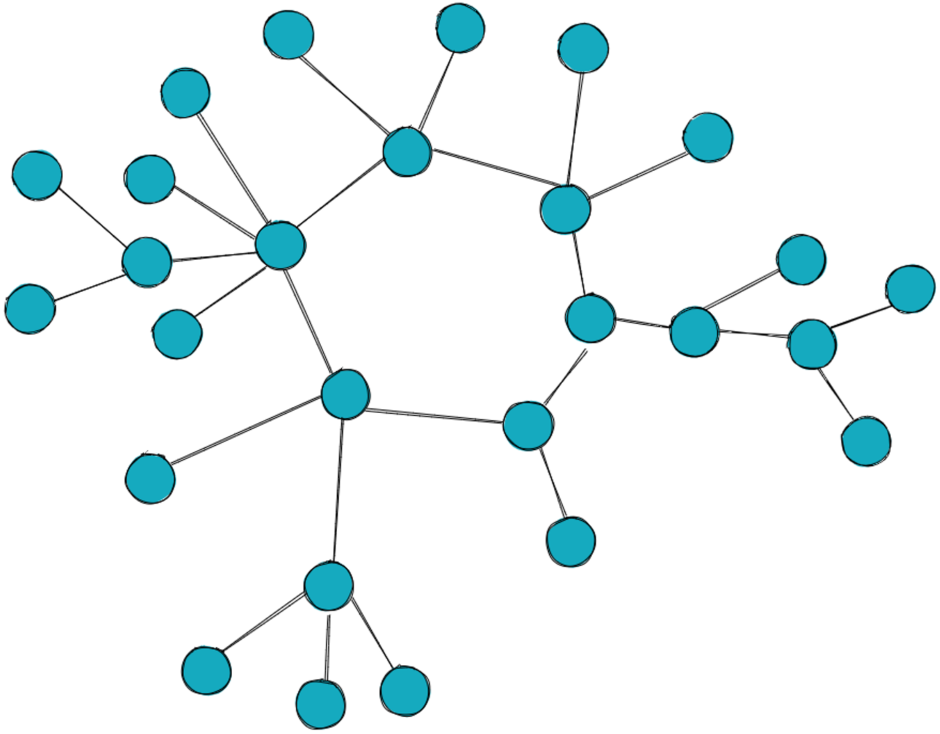
3. re-route it safely



# Biomedical knowledge is spread across multiple resources


depmap
BioGrid
Connecting health information
uberon
HGNC
Cellosaurus
PubMed
GOSTAR
Global Online Structure Activity Relationship Database
Ensembl
MedDRA
OMIM
Human Genetics Knowledge for the World
EBI
Expression Atlas
ChEMBL
Open Targets
mondo
THE WORLD'S DISEASE CONCEPTS, UNIFIED
DisGeNET

# Graph makes things simpler

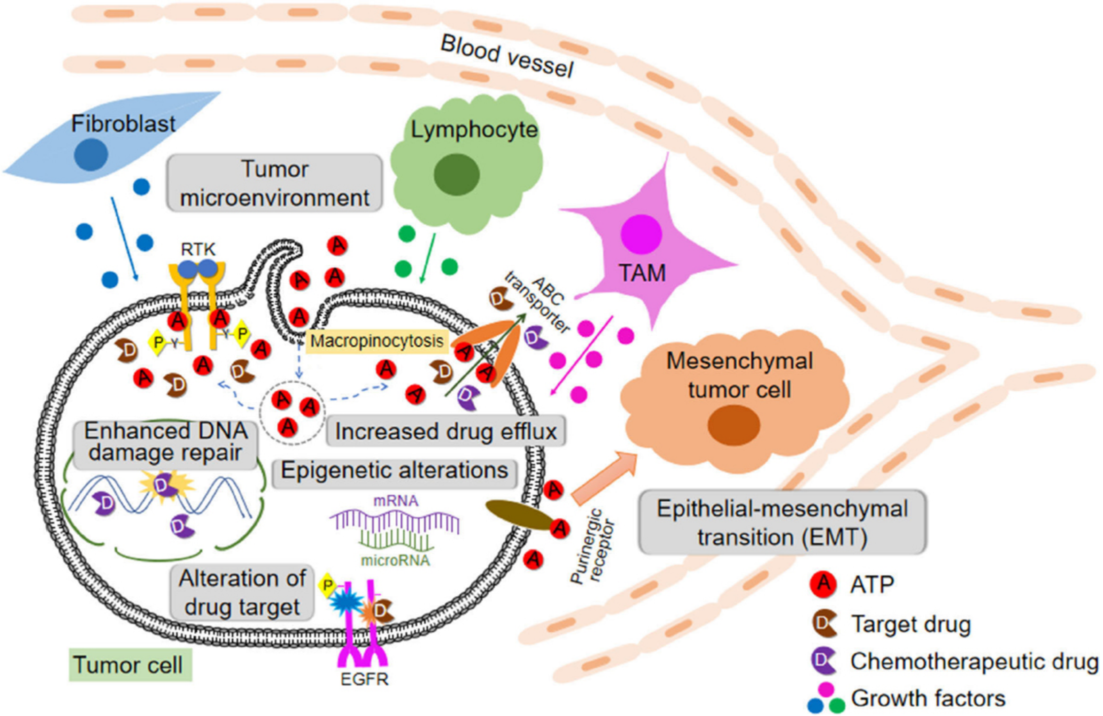


- Biomedical information often comes in forms of **networks** and **hierarchies**
- Graph is a convenient way to organise it
- BIKG (our internal knowledge graph): **60+** data sources including - omics and data extracted from the literature
- **11 M nodes, 1 B edges**
- Use graph as a source of context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

# Early success story:

# graph-based recommendations

# Applied recommendation problem #1: contextualize experimental data

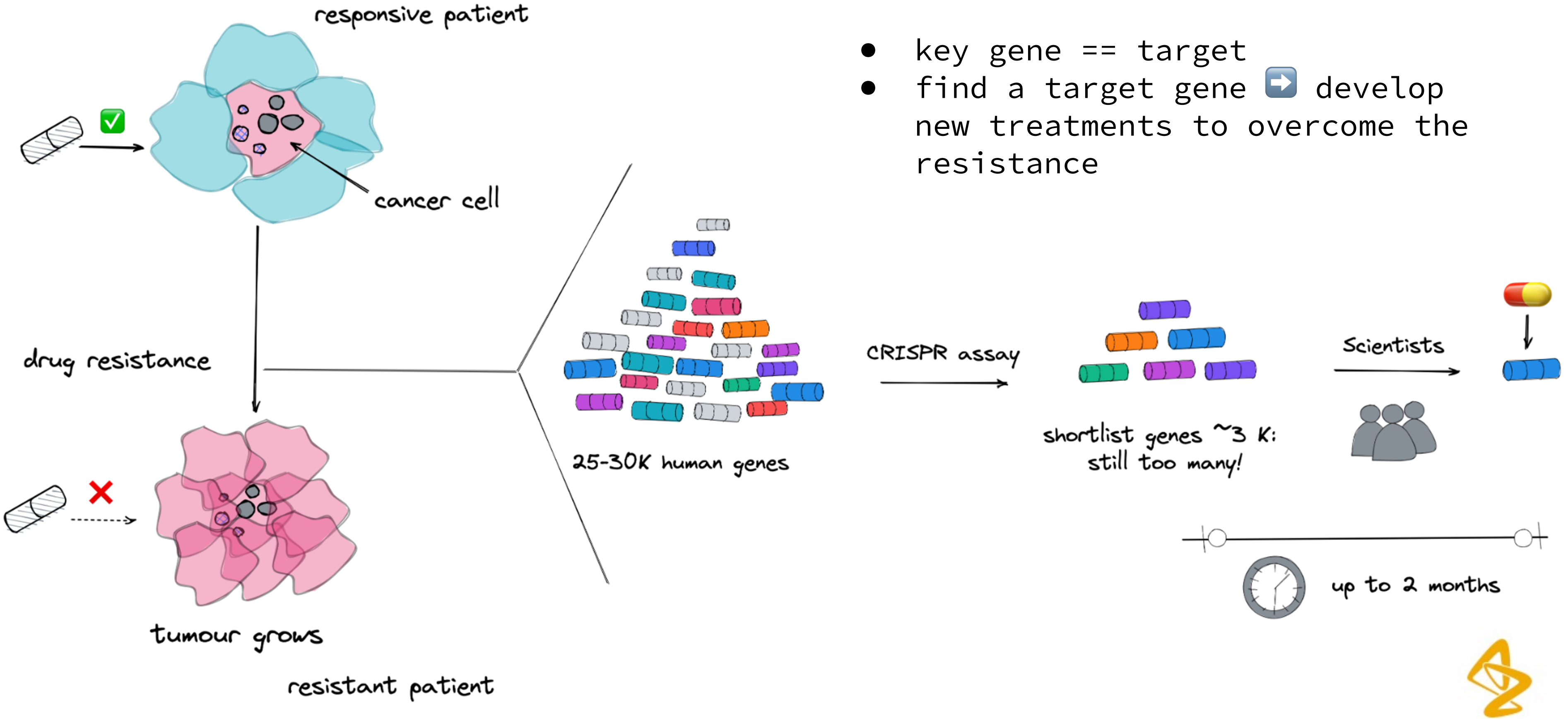


X Wang, H Zhang, X Chen - *Cancer Drug Resistance, 2019*

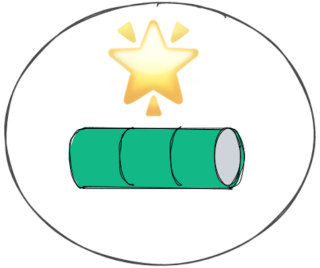
- Drug resistance in lung cancer
- Occurs in a sub-population of patients
- Resistance landscape is complex

# How to help scientist find key genes faster?

- key gene == target
- find a target gene ➡️ develop new treatments to overcome the resistance



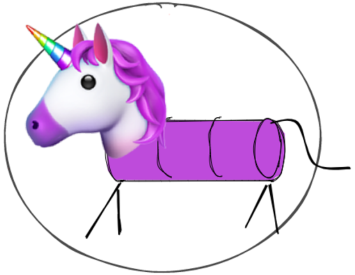
# An ideal target



- ☑ 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

...

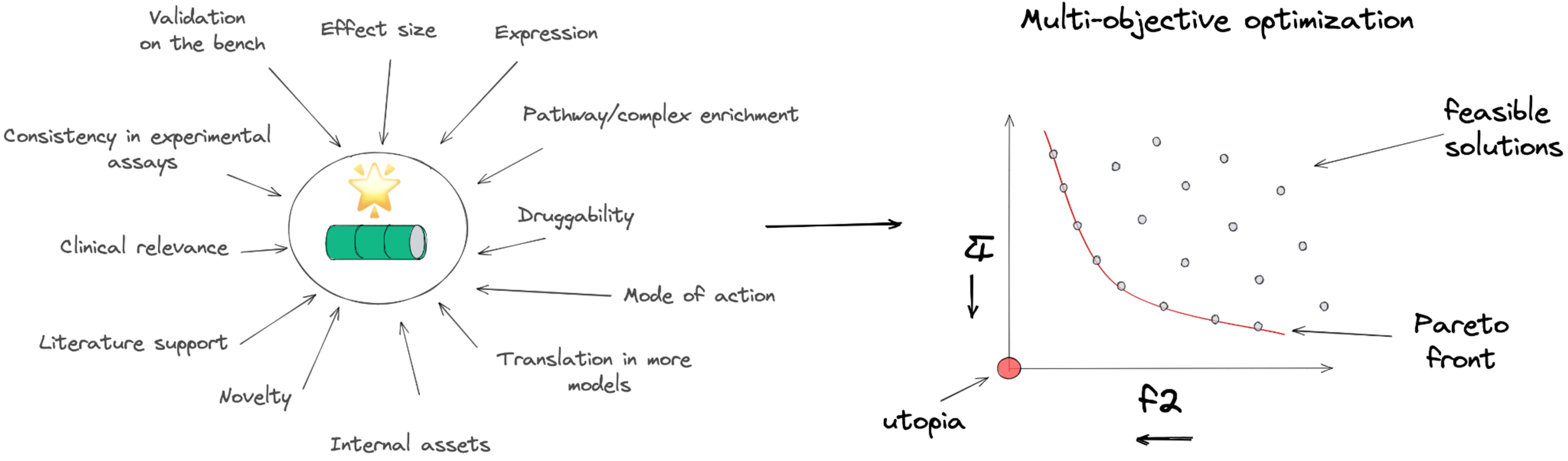
# An ideal target does not exist



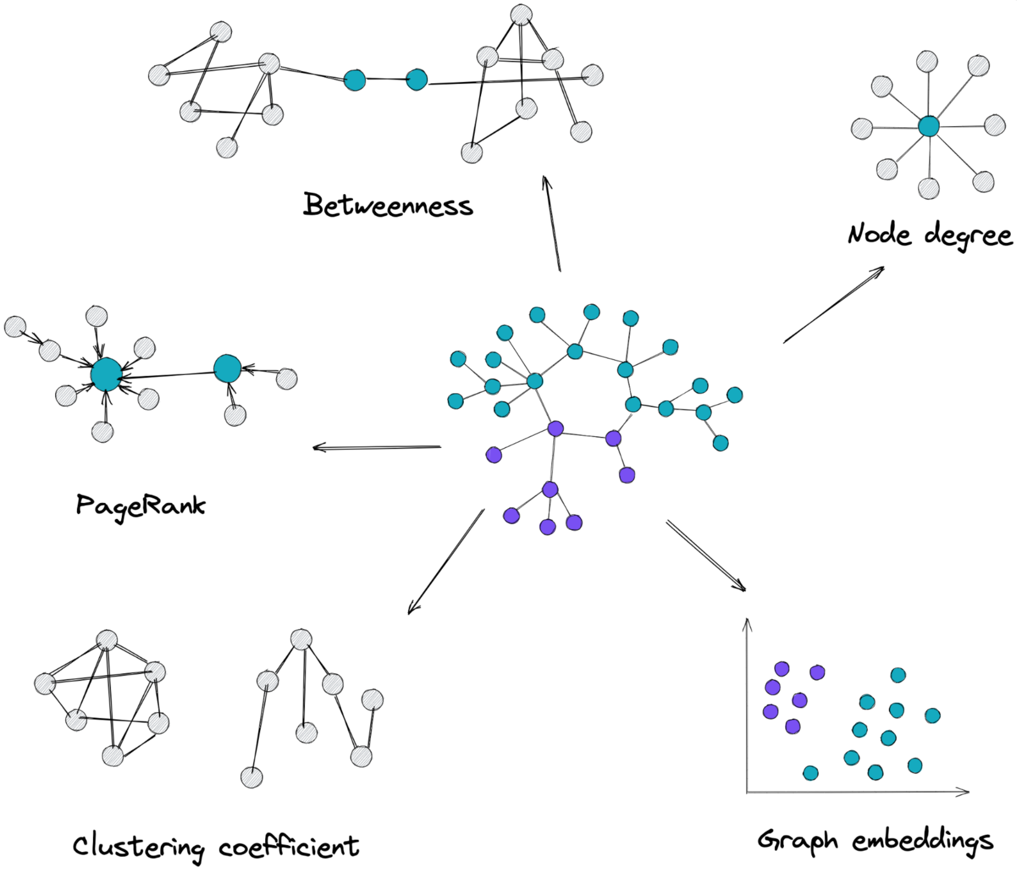
- ☑ Expression
- ☑ Pathway/complex enrichment
- ☑ Effect size
- ☑ Druggability
- ☑ Mode of action
- ☑ Translation in models
- ☑ Internal assets
- ☑ Bench validation
- ☑ Consistency in assays
- ☑ Clinical relevance
- ☑ Literature support
- ☑ Novelty

...

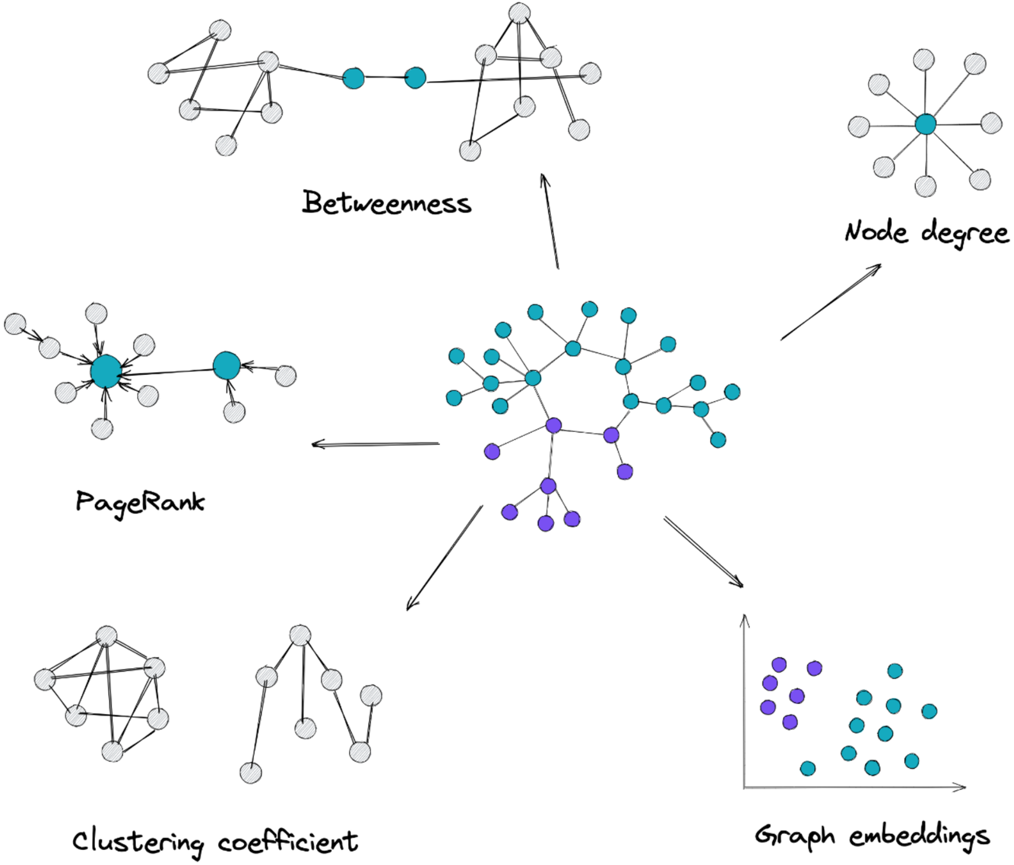
# Target selection as an optimization problem



# Hybrid feature set: source features from the graph

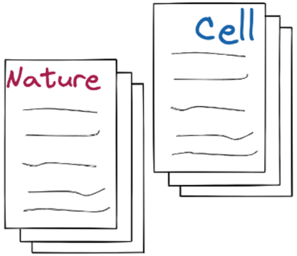


# Hybrid feature set: combine with clinical features



Clinical features

Literature support

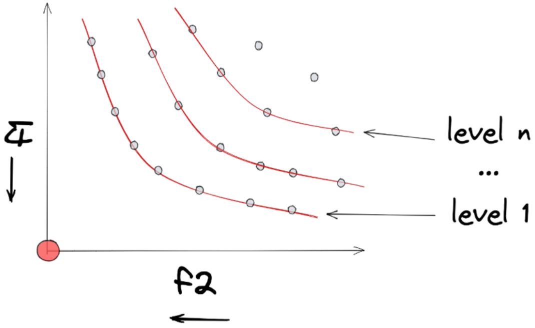


Druggability

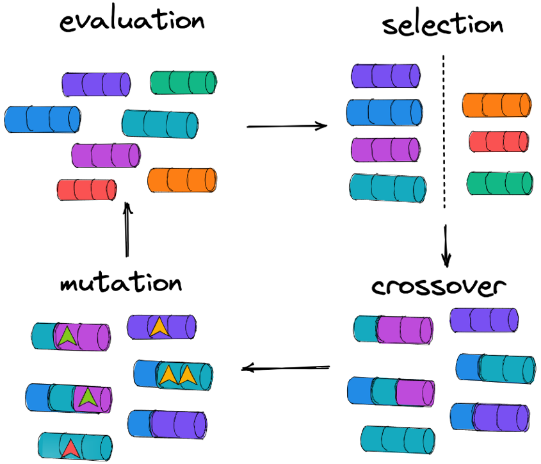
Pre-clinical experimental assays

# Approaches

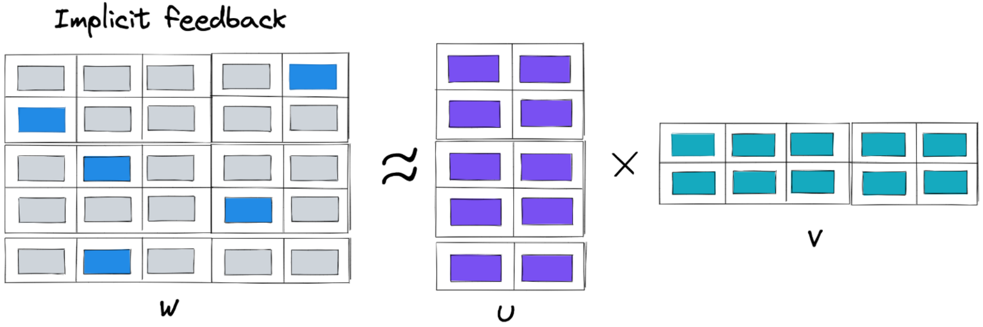
## 1 Compute exact Pareto front



## 2 Evolutionary algorithms

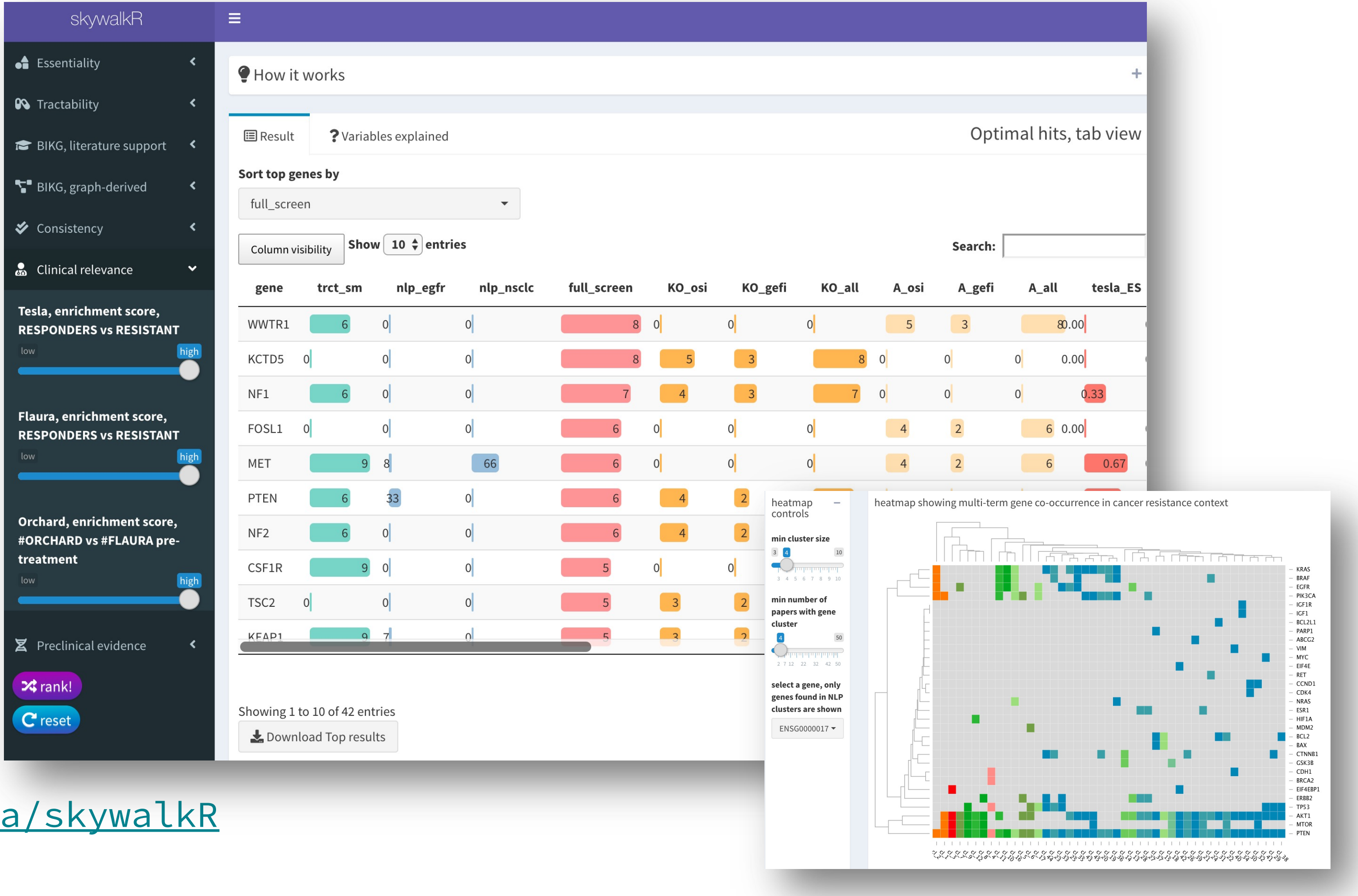


## 3 Matrix factorization



# SkywalkR, interactive interface

- select a subset of objectives
- set optimization directions
- explore trade-offs


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:
Download Top results
heatmap controls
min cluster size
3 4 10
min number of papers with gene cluster
4 50
select a gene, only genes found in NLP clusters are shown
ENSG0000017
heatmap showing multi-term gene co-occurrence in cancer resistance context
KRAS
BRAF
EGFR
PIK3CA
IGF1R
IGF1
BCL2L1
PARP1
ABCG2
VIM
MYC
EIF4E
RET
CCND1
CDK4
NRAS
ESR1
HIF1A
MDM2
BCL2
BAX
CTNNB1
GSK3B
CDH1
BRCA2
EIF4EBP1
ERBB2
TP53
AKT1
MTOR
PTEN
cl_2 cl_1 cl_3 cl_7 cl_9 cl_12 cl_8 cl_4 cl_11 cl_10 cl_16 cl_5 cl_6 cl_17 cl_44 cl_23 cl_33 cl_25 cl_35 cl_43 cl_45 cl_20 cl_19 cl_36 cl_14 cl_13 cl_28 cl_27 cl_37 cl_15 cl_18 cl_42 cl_26 cl_39 cl_21 cl_24 cl_31 cl_22 cl_40 cl_34 cl_30 cl_32 cl_41 cl_29 cl_38

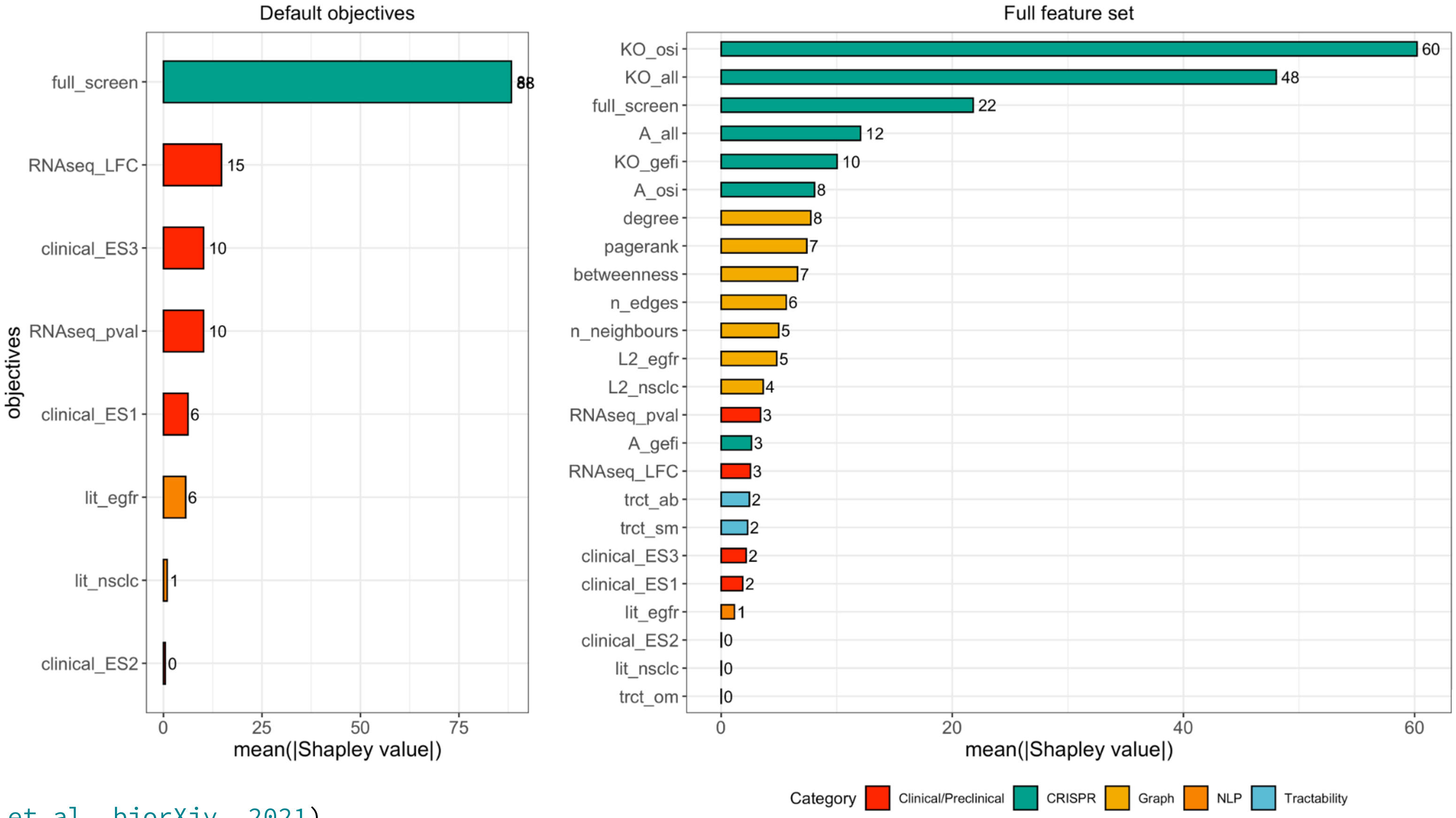
| gene | trct_sm | nlp_egfr | nlp_nsclc | full_screen | KO_osi | KO_gefi | KO_all | A_osi | A_gefi | A_all | tesla_ES |
|---|---|---|---|---|---|---|---|---|---|---|---|
| WWTR1 | 6 | 0 | 0 | 8 | 0 | 0 | 0 | 5 | 3 | 8 | 0.00 |
| KCTD5 | 0 | 0 | 0 | 8 | 5 | 3 | 8 | 0 | 0 | 0 | 0.00 |
| NF1 | 6 | 0 | 0 | 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 | 5 | 3 | 2 | | | | | |

Showing 1 to 10 of 42 entries

github.com/AstraZeneca/skywalkR

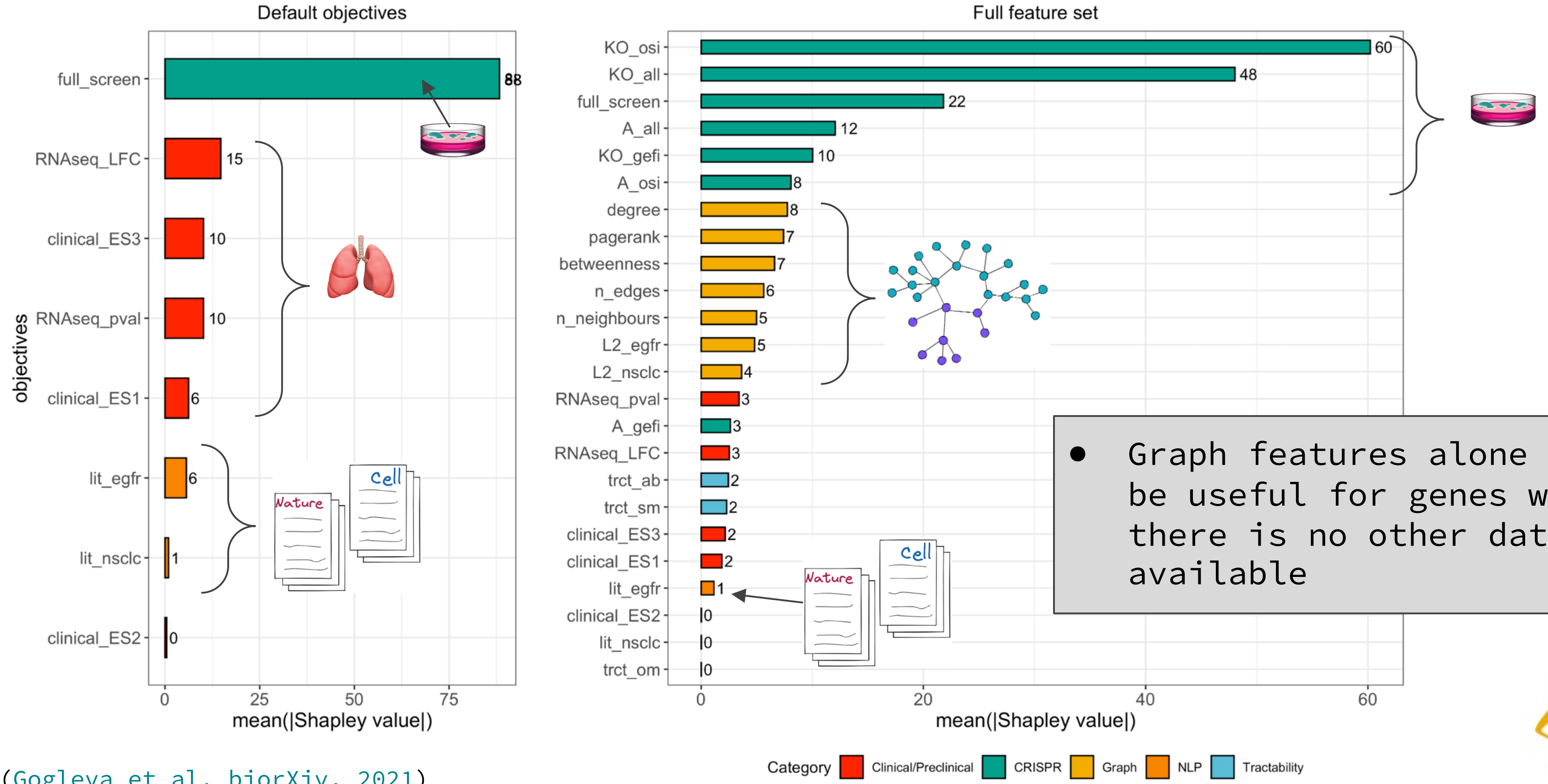
# Imperfect validation

# Model domain scientist as a black box classifier



(Gogleva et al, biorXiv, 2021)

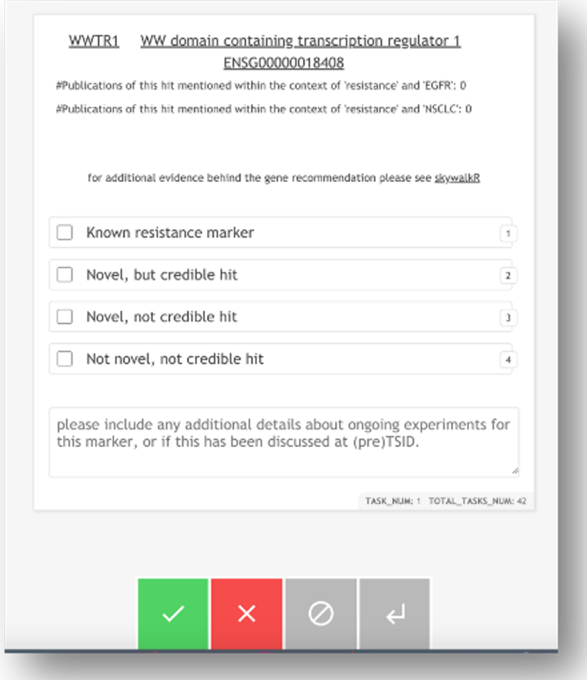
# Graph-derived features follow clinical in unbiased setting



- Graph features alone can be useful for genes when there is no other data available

(Gogleva et al, biorXiv, 2021)

# 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

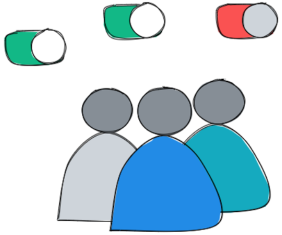
#Publications of this hit mentioned within the context of 'resistance' and 'NSCLC': 0

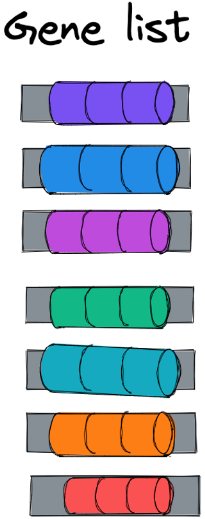
for additional evidence behind the gene recommendation please see skywalkR

- □ 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.

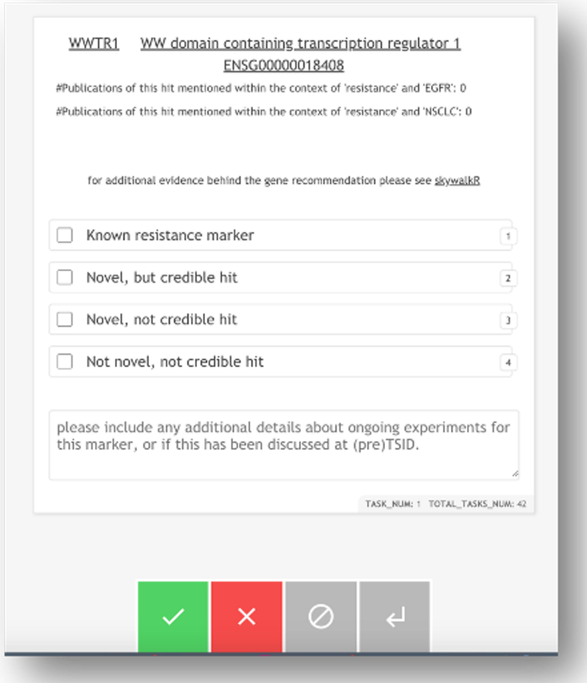
TASK_NUM: 1 TOTAL_TASKS_NUM: 42





prodigy

# 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

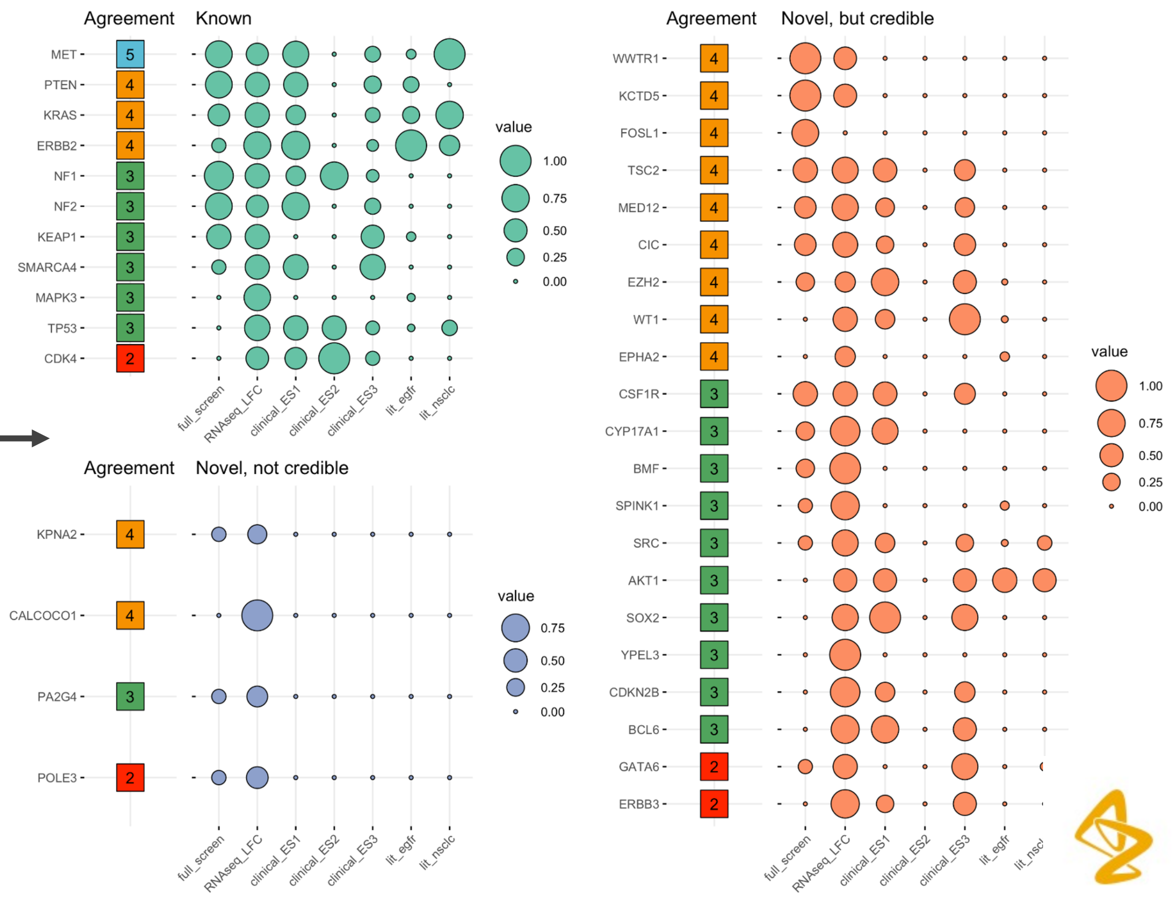
#Publications of this hit mentioned within the context of 'resistance' and 'NSCLC': 0

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

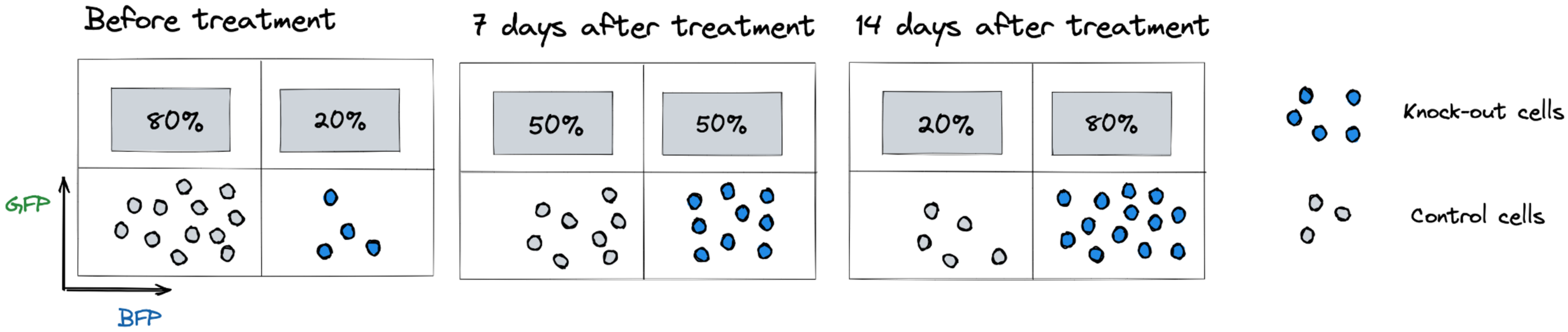
please include any additional details about ongoing experiments for this marker, or if this has been discussed at (pre)TSID.

TASK_NUM: 1 TOTAL_TASKS_NUM: 42



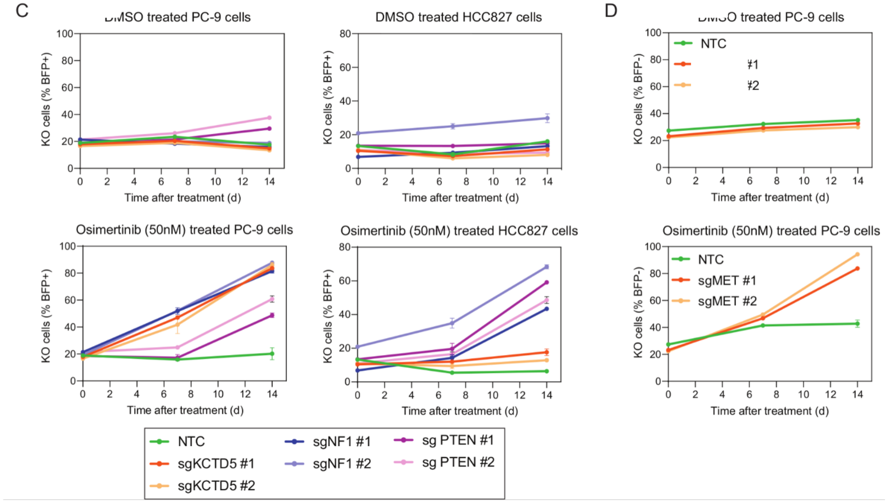
(Gogleva et al, biorXiv, 2021)

# Experimental validation *in vitro*



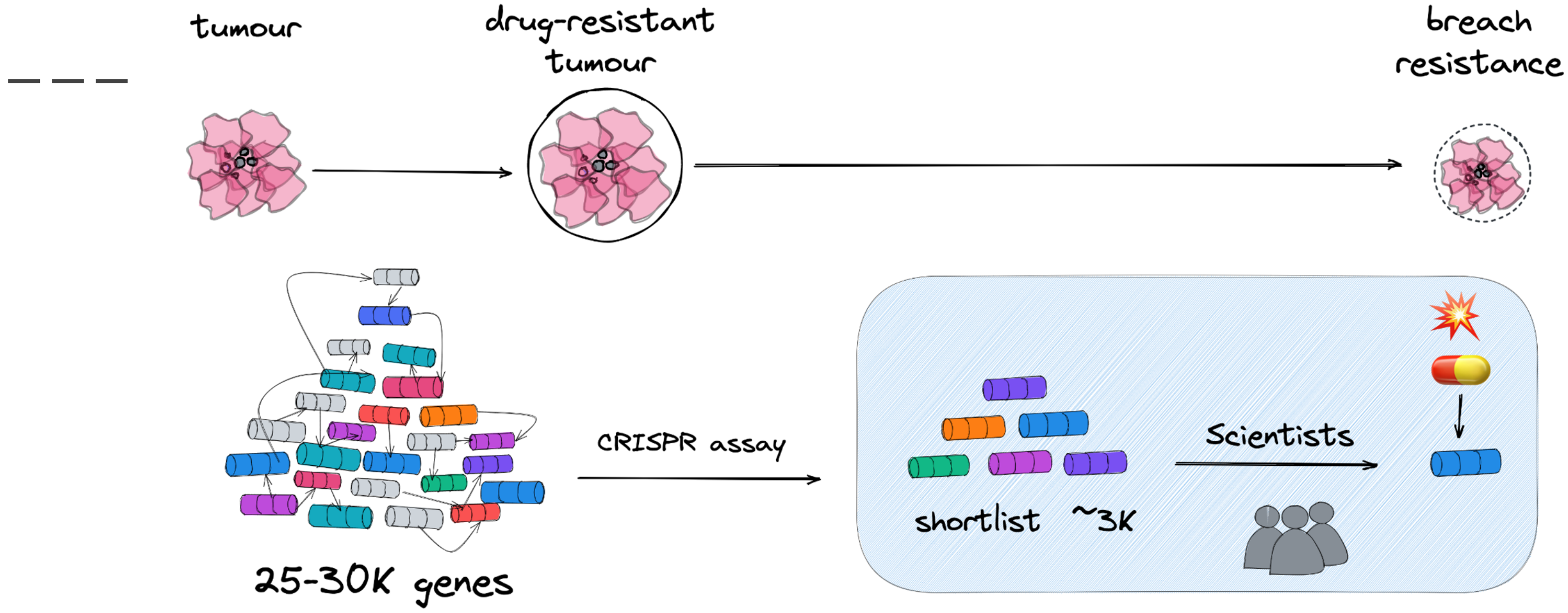
- confirmed involvement of 4 recommended genes in drug resistance

- next: test the remaining genes



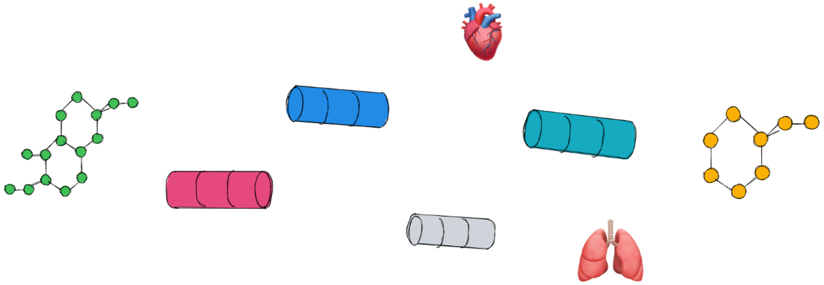
(Gogleva et al, biorXiv, 2021)

# Imperfect, yet already useful recommendation system



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

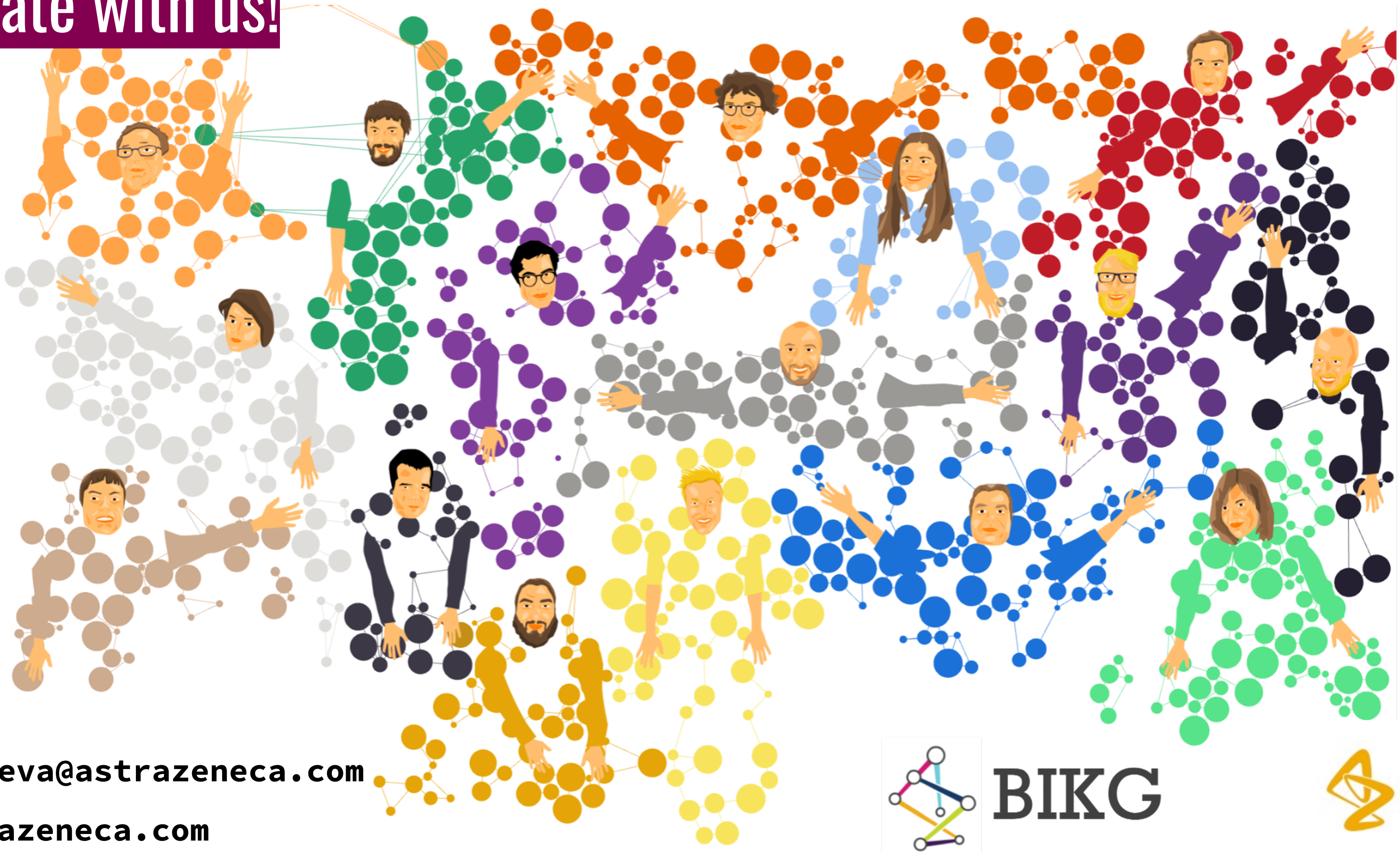
# Take home message



- 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/recsys21gogleva/**

# Collaborate with us!



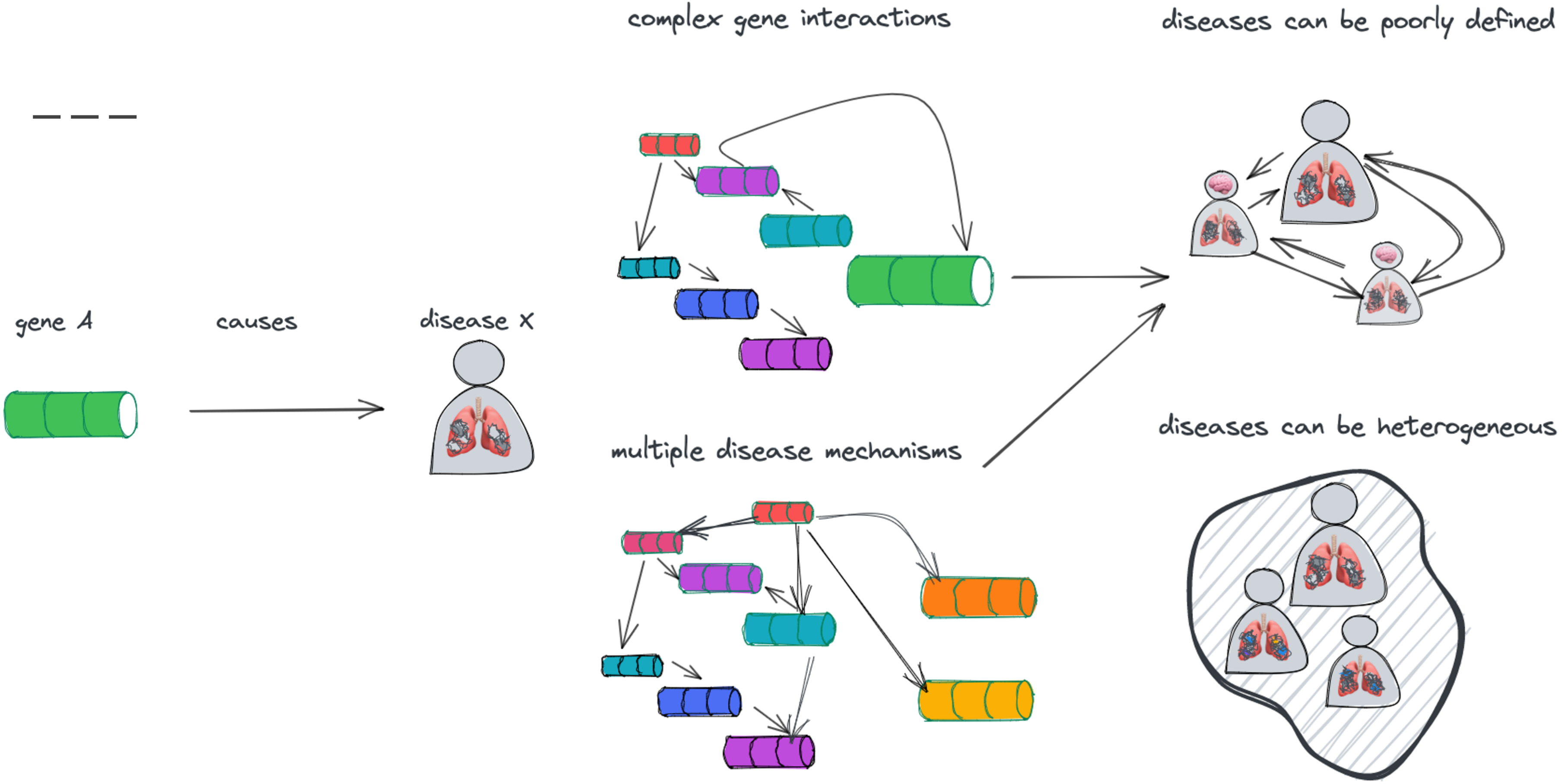
anna.gogleva@astrazeneca.com

bikg@astrazeneca.com

BIKG

# Translating recommendation approaches to biomedical field: a few complications

# Biological entities are complex



gene A

causes

disease X

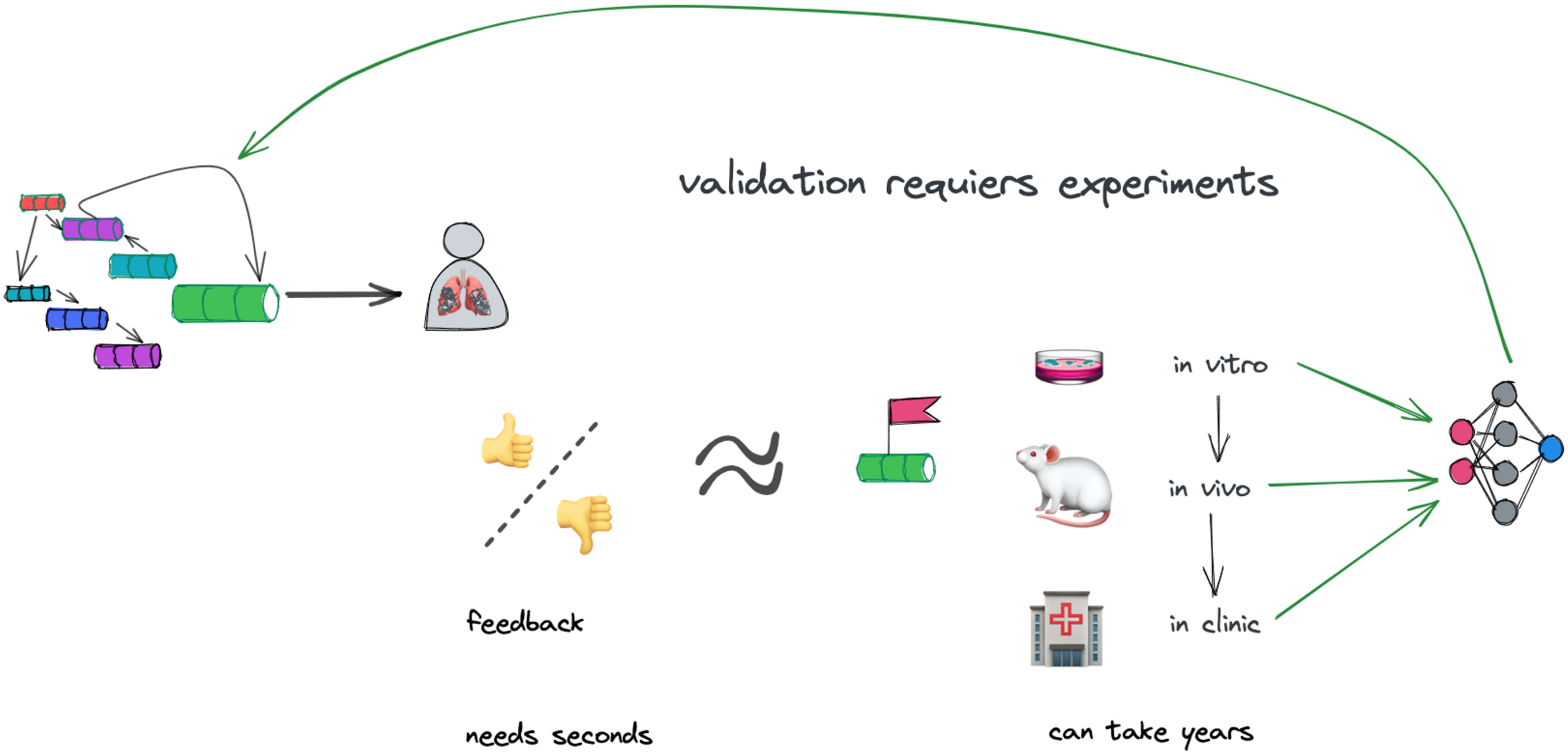
complex gene interactions

multiple disease mechanisms

diseases can be poorly defined

diseases can be heterogeneous

# Validation is slow and expensive



validation requiers experiments

in vitro

in vivo

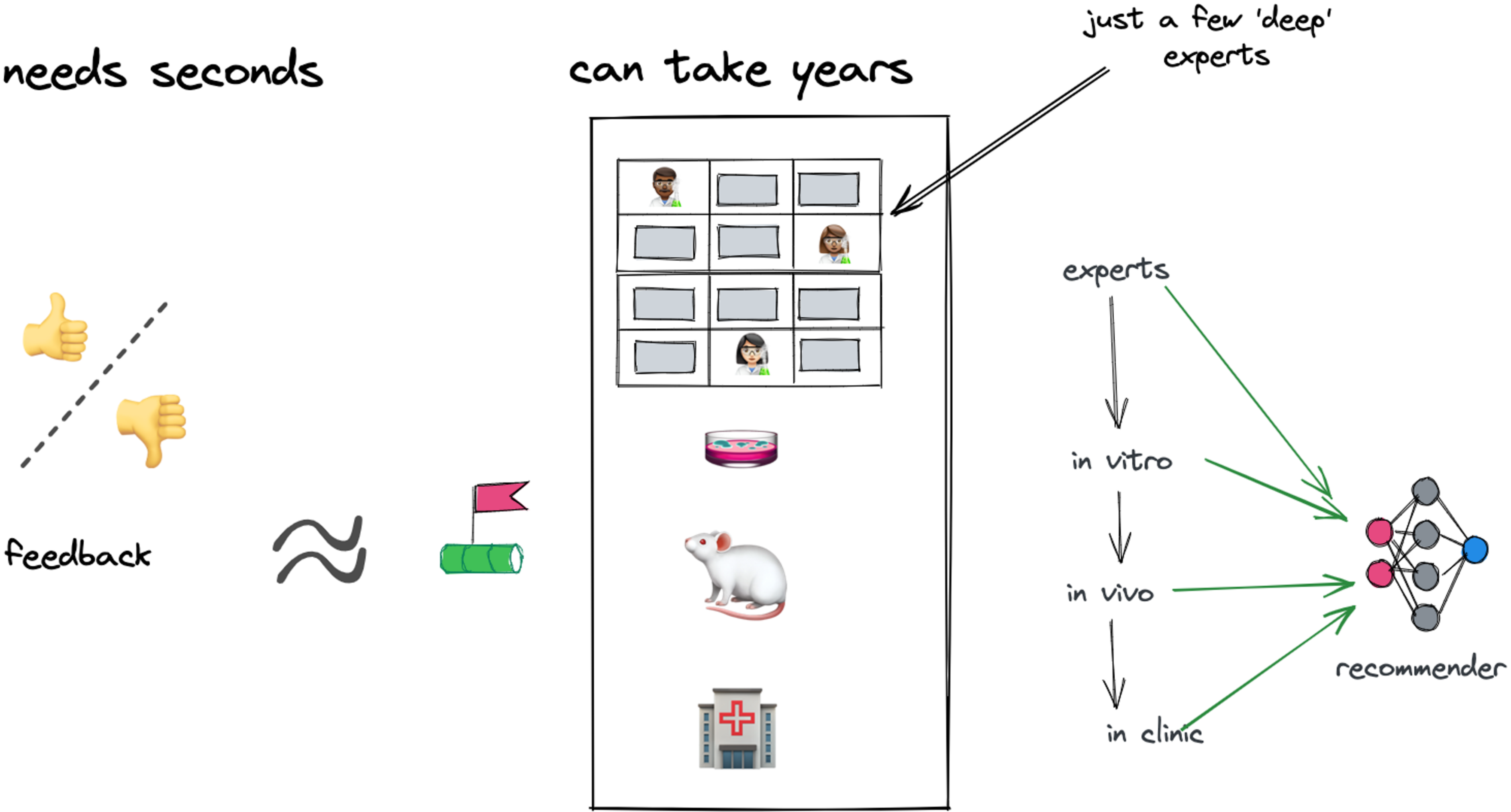
in clinic

feedback

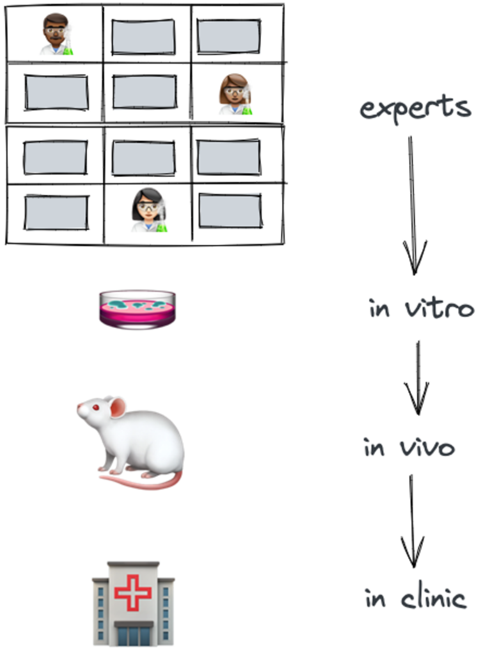
needs seconds

can take years

# Implicit & explicit feedback is scarce



# Team of experts rather than a single user makes decisions



experts

in vitro

in vivo

in clinic

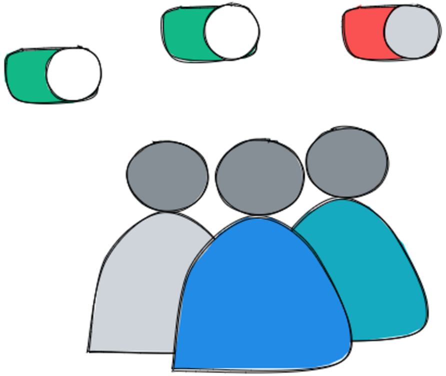
feedback

needs seconds

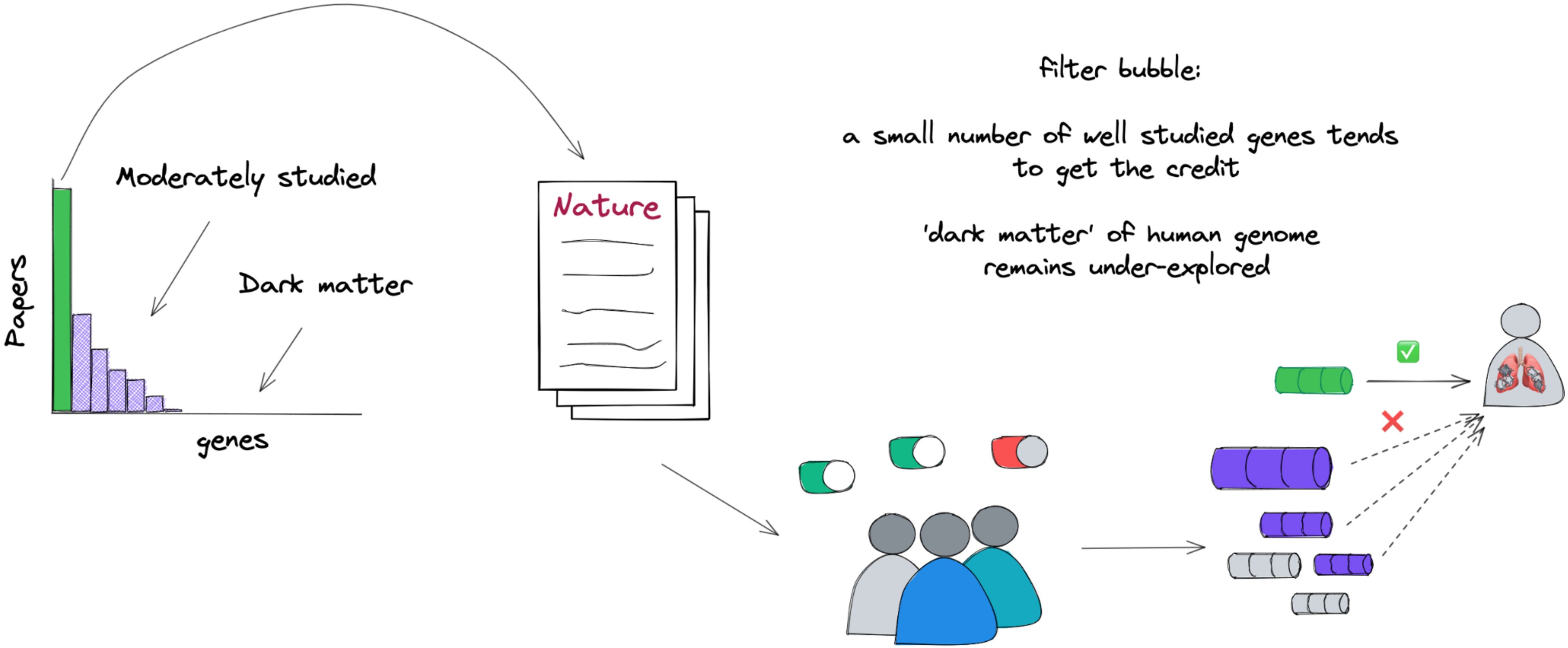
can take years

a team of experts makes decisions

each expert has their own bias



# Previous literature biases users decisions

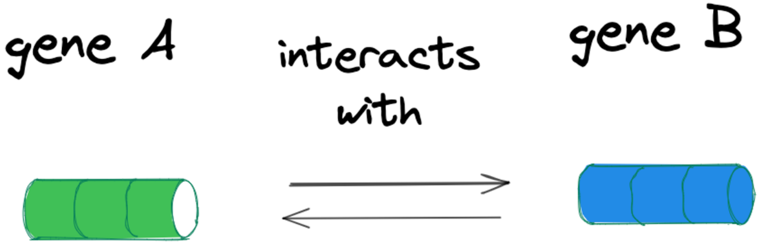


filter bubble:

a small number of well studied genes tends to get the credit

'dark matter' of human genome remains under-explored

# Ground truths are rare  and context-specific



gene A interacts with gene B

⏱️ same time:
- disease stage
- developmental stage

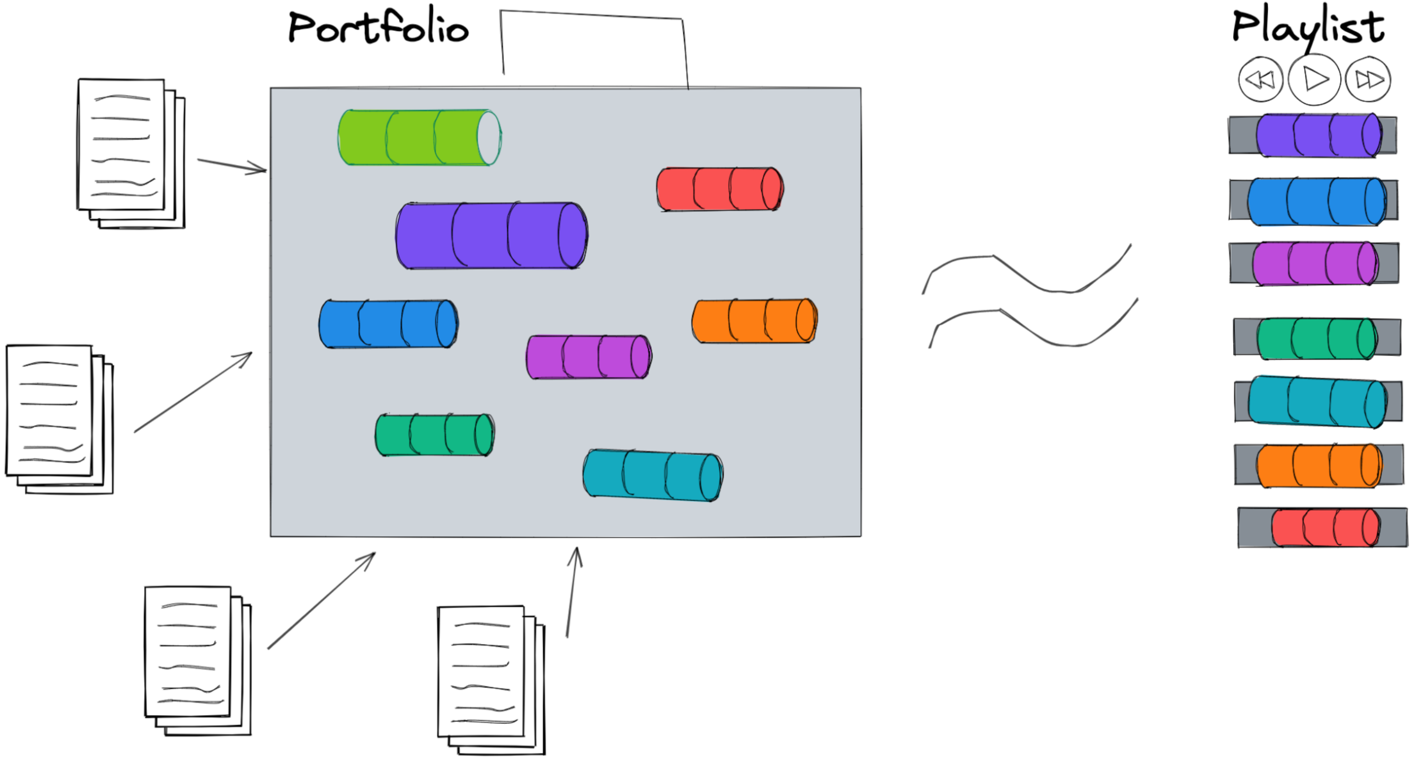
📍 same place:
- tissue
- organ

🧬 same genetic background

😭 there is a lot of data out there,
but never the data you need to train your model

# Portfolio problem vs single choice:

## continuously optimize based on constantly changing evidence

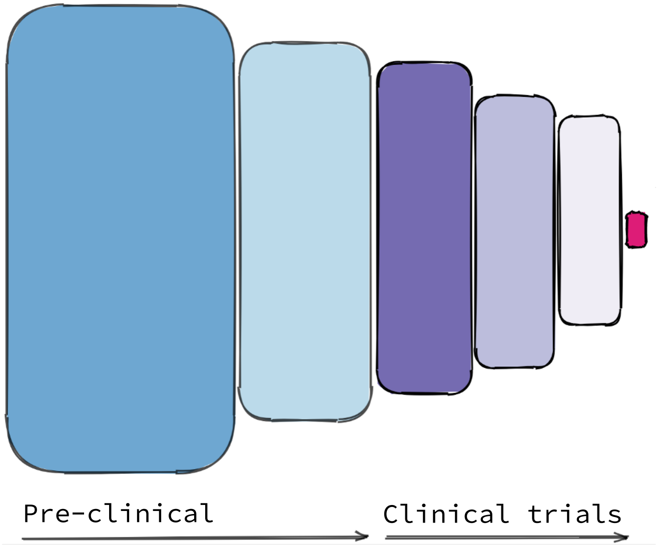


# Supplementary:
supervised recommendations

# Can we learn from previous drug trials?

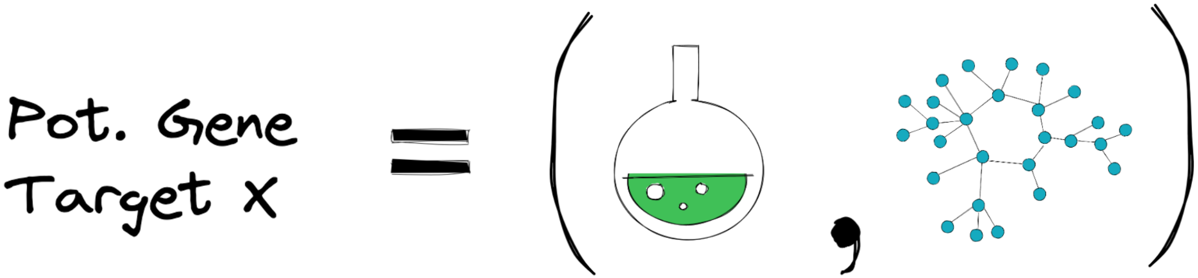
- Thousands of clinical trials preclinical experiments (internal + external)

- Idea: use data on previous (potential) targets as training data for a supervised model



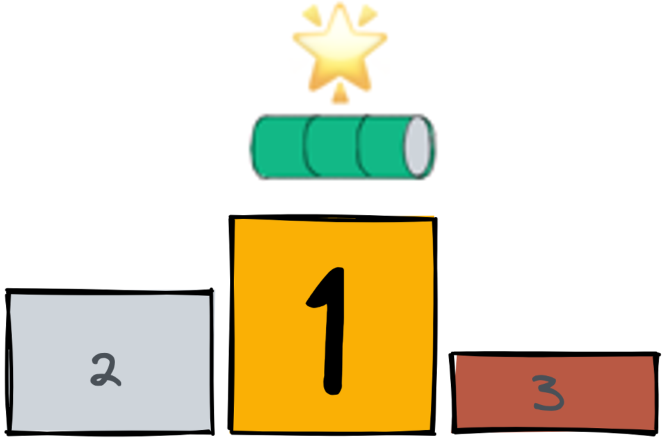
# Can we learn from previous drug trials?

- Represent genes with experimentally derived and KG-derived features
    - Experimental - activity in certain bio processes
    - KG-derived - graph distances, embedding distances, etc. etc.



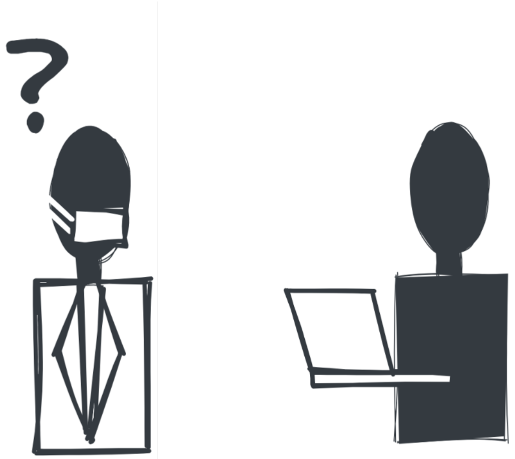
# Can we learn from previous drug trials?

- Train a supervised ranking model (LightGBM) with randomly sampled targets as negatives and clinically promising targets as positives



# Human-Model trust

- We need biologist’s to sign off on our model’s recommendations
- For that, we need their trust
  - NDCG or other “ML” metrics mean nothing to a biologist
  - Biologists expect certain genes as a sanity-check



# Human-Model trust

“I would expect to see Gene X in your recommendations - otherwise we have a problem”

# Human-Model trust

“Yup the model is recognizing Gene X as a promising gene target!”

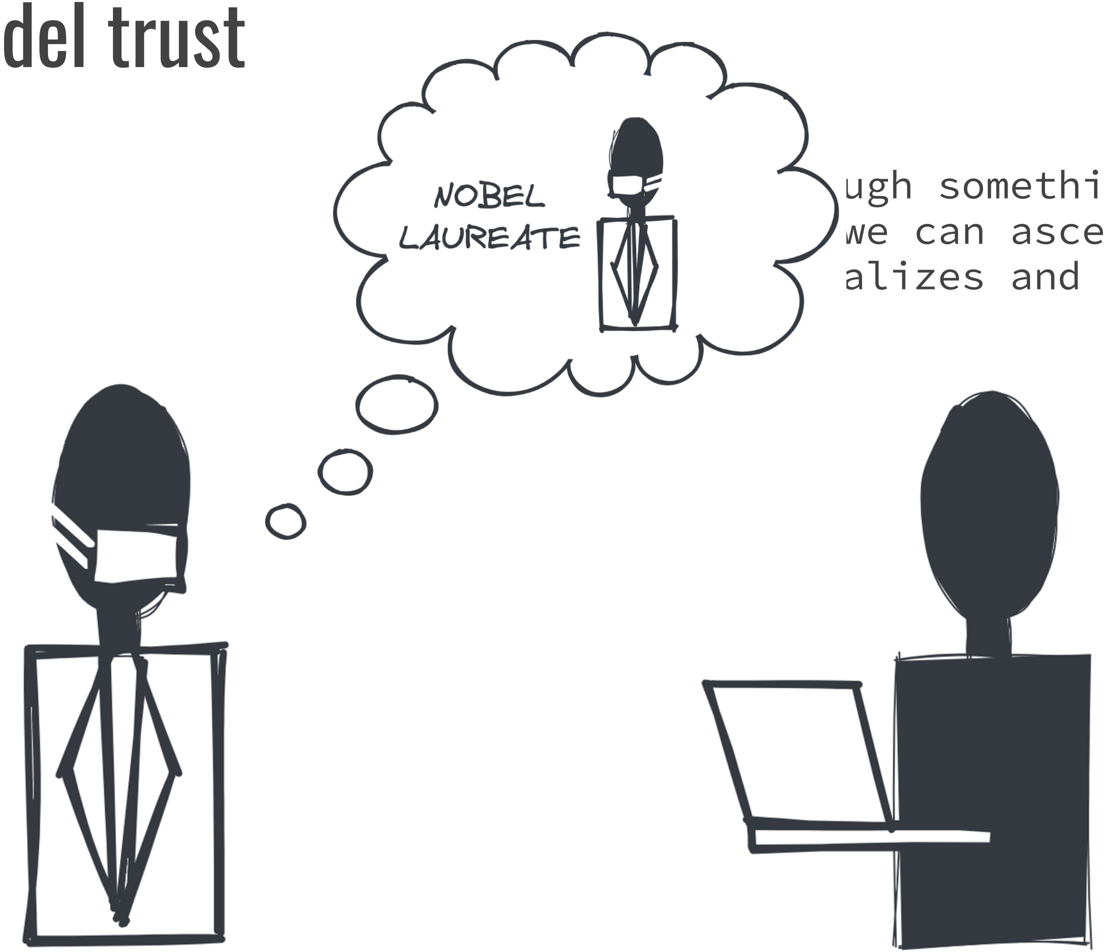
# Human-Model trust

“How do I know the model isn’t just regurgitating what I’ve told you?”

# Human-Model trust

“Well, through something called cross validation we can ascertain that the model generalizes and –”

# Human-Model trust



ugh something called cross
we can ascertain that the
alizes and –”

# Human-Model trust

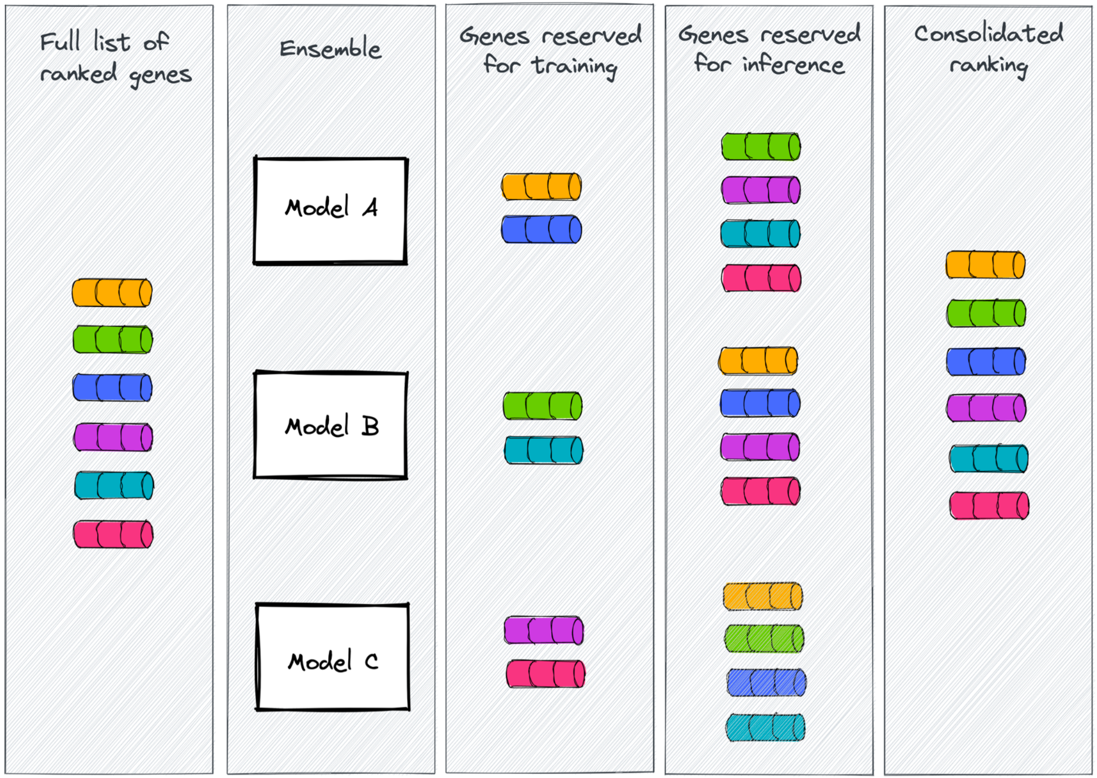
- Problem: human genome is finite
  (Since we rank the full genome, the training set will exist somewhere in the final model output)

- How can we guarantee that no “regurgitation” is happening during inference?

# “Honest” Ensembling

- Training data is split among an ensemble of models
- If a gene has been seen by a model during training - this model can’t rank its target-aptitude during inference

# “Honest” Ensembling


Full list of ranked genes
Ensemble
Model A
Model B
Model C
Genes reserved for training
Genes reserved for inference
Consolidated ranking

# Jury is still out

- Training data: genes that have previously been found promising in COPD (Chronic obstructive pulmonary disease)
- After ranking:
  - Take the top ~200 genes
  - Filter for known involvement in a number of interesting molecular processes
  - Bring to biologists for manual quality control
- => 29 potential gene targets are now in experimental validation