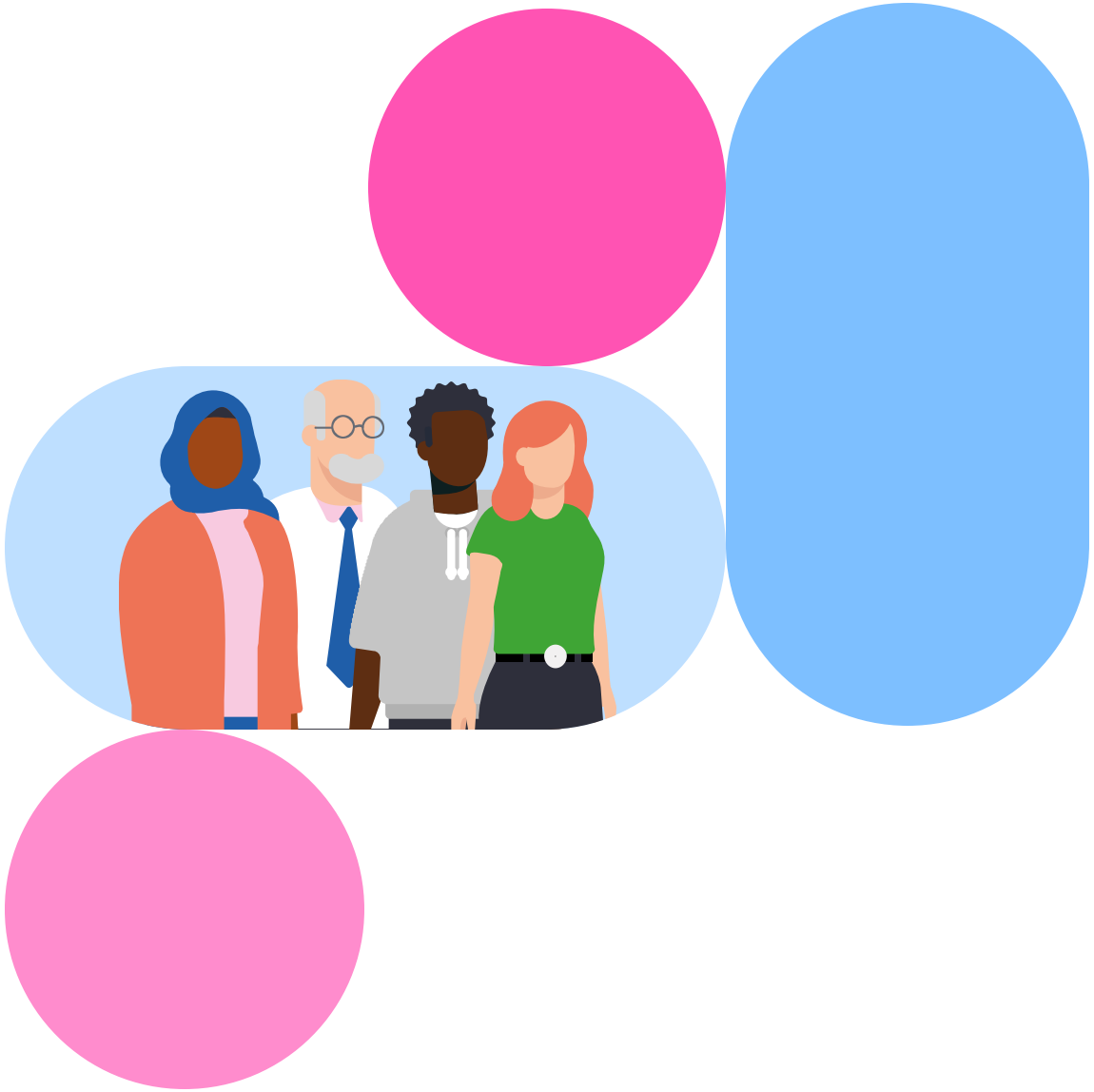
A demonstration of accelerating scientific literature searches with machine learning

Samuel Barnett, Eleanor Williams and Emily Perry



24th June 2025

Introducing ourselves

Samuel Barnett

- Wet-lab molecular biologist for 10 years
- Machine learning Engineer at Genomics England for 2 years
- Technical lead in the Data Enrichment squad improving researcher data

Eleanor Williams

- Been a curator (gene expression, imaging, gene-disease data) for 20 years
- Worked at Genomics England for 8 years
- Leads the team working on gene panels for the National Genomic Test Directory

Emily Perry

- 13 years teaching and training in bioinformatics and genomics
- Four years with Genomics England
- Runs the training programme for researchers using the Genomics England Research Environment

Learning outcomes

Understand the **use case** for using machine learning in biocuration

Appreciate the **challenges** in finding evidence for gene-disease associations in literature

Be familiar with the **core workflow** using Machine Learning to search for relevant publications for curation

Be able to **recognise the advantages and disadvantages** of using machine learning tools for literature searching

Agenda

Why are we searching for literature at Genomics England?

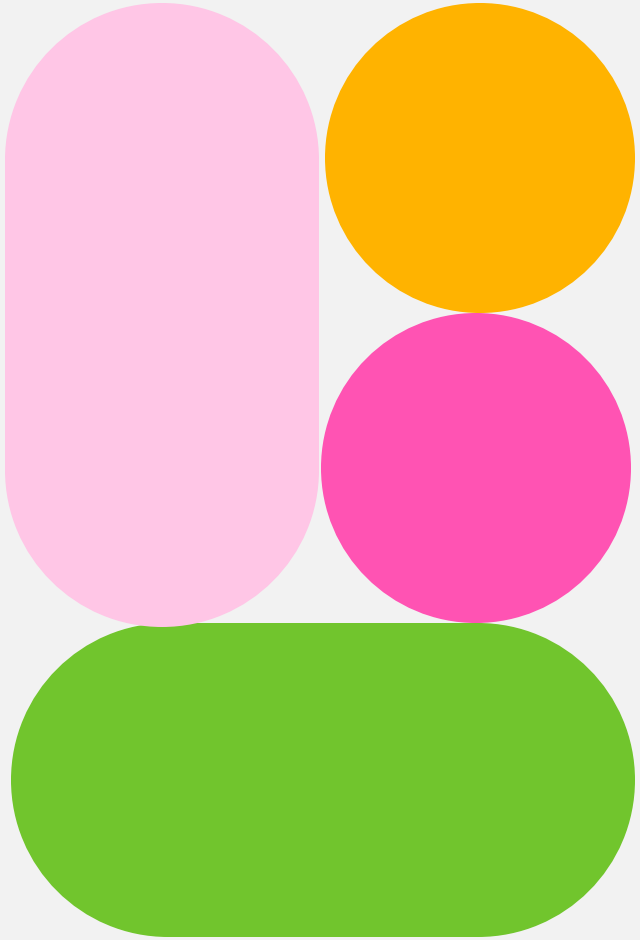
Hands on – why is searching for literature difficult?

Machine Learning and text processing – intro to our trial and demo

Q & A

ML vs human – head to head comparison

Was our ML experiment for biocuration sucessfull?



Why are we searching for literature at Genomics England?

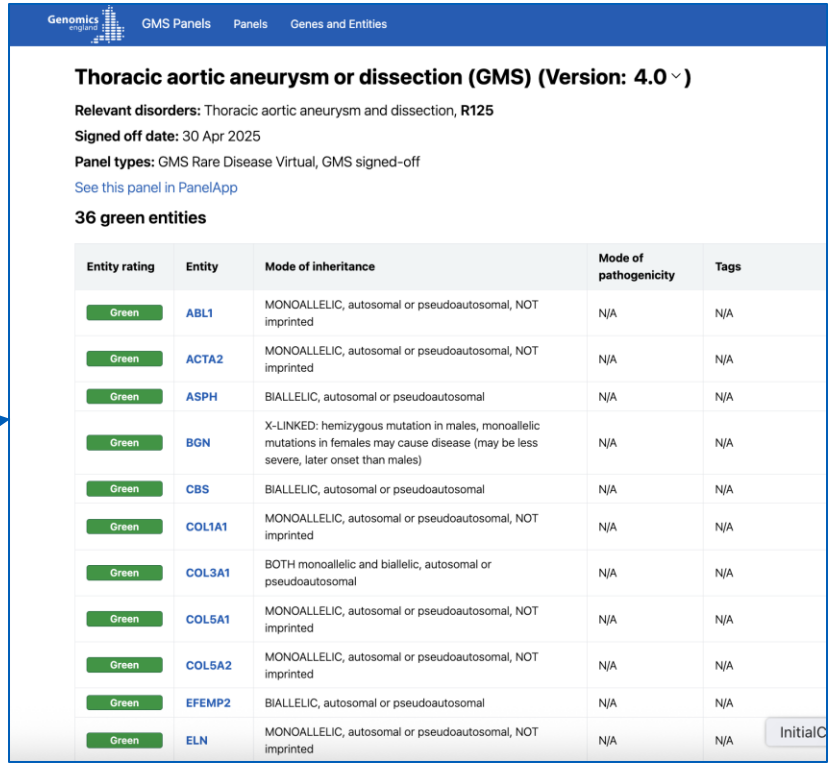
Genomics England and the Genomic Medicine Service

NHS

National Genomic Test Directory

Content

https://www.england.nhs.uk/publication/national- genomic-test-directories



https://nhsgms-panelapp.genomicsengland.co.uk/
https://panelapp.genomicsengland.co.uk/

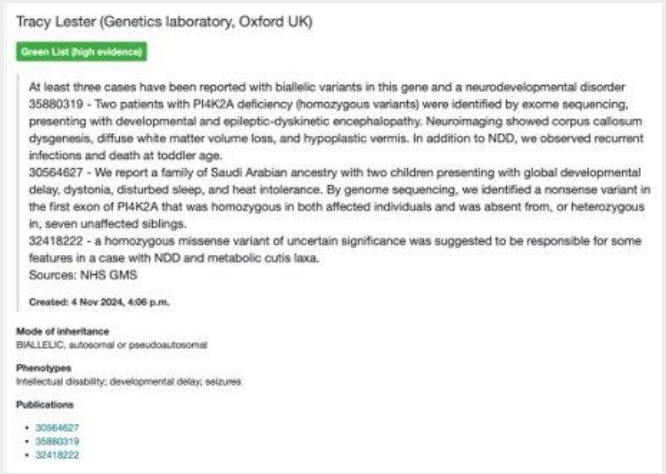
> 300 GMS panels

Gene panels

PanelApp

https://panelapp.genomicsengland.co.uk

PanelApp is an open platform which captures evidence for gene-disease relationships



Tracy Lester (Genetics laboratory, Oxford UK)

Green List (high evidence)

At least three cases have been reported with biallelic variants in this gene and a neurodevelopmental disorder
35880319 - Two patients with PI4K2A deficiency (homozygous variants) were identified by exome sequencing, presenting with developmental and epileptic-dyskinetic encephalopathy. Neuroimaging showed corpus callosum dysgenesis, diffuse white matter volume loss, and hypoplastic vermis. In addition to NDD, we observed recurrent infections and death at toddler age.
30564627 - We report a family of Saudi Arabian ancestry with two children presenting with global developmental delay, dystonia, disturbed sleep, and heat intolerance. By genome sequencing, we identified a nonsense variant in the first exon of PI4K2A that was homozygous in both affected individuals and was absent from, or heterozygous in, seven unaffected siblings.
32418222 - a homozygous missense variant of uncertain significance was suggested to be responsible for some features in a case with NDD and metabolic cutis laxa.
Sources: NHS GMS

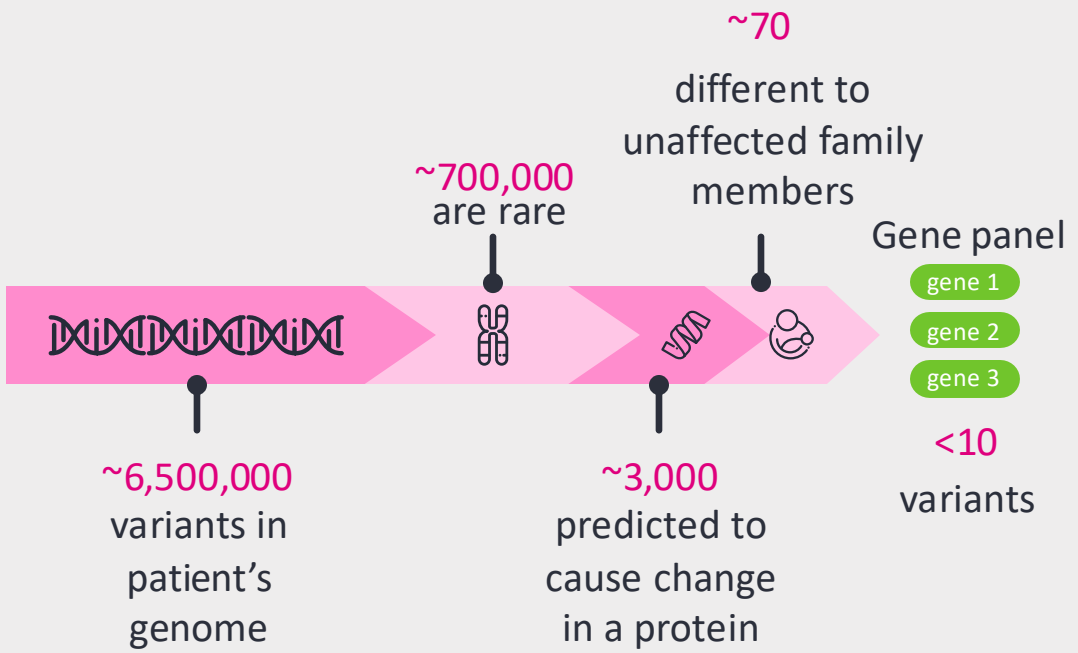
Created: 4 Nov 2024, 4:06 p.m.

Mode of inheritance
BIALLELIC, autosomal or pseudoautosomal

Phenotypes
Intellectual disability; developmental delay; seizures

Publications
- 30564627
- 35880319
- 32418222

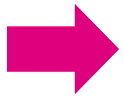
Gene panels used to prioritise variants for clinical scientist to look at



New evidence applied to clinical practise

Panel v1

- ACO2
- GJC2
- POMK
- UPB1
- NAV3
- MYO5B



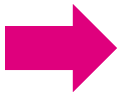
Hum Genet. 2025 Jan;144(1):55-65. doi: 10.1007/s00439-024-02718-6. Epub 2024 Dec 21.

Further evidence of biallelic NAV3 variants associated with recessive neurodevelopmental disorder with dysmorphism, developmental delay, intellectual disability, and behavioral abnormalities

Naseebullah Kakar [1 2], Selinda Mascarenhas [# 3], Asmat Ali [# 4], Azmatullah [# 5], Syed M Ijlal Haider [6], Vaishnavi Ashok Badiger [3], Mobina Shadman Ghofrani [1], Nathalie Kruse [1], Sohana Nadeem Hashmi [4], Jelena Pozojevic [1], Saranya Balachandran [1], Mathias Toft [7 8], Sajid Malik [5], Kristian Händler [1], Ambrin Fatima [4], Zafar Iqbal [8], Anju Shukla [3], Malte Spielmann [9], Periyasamy Radhakrishnan [10]

Affiliations + expand

PMID: 39708122 PMCID: PMC11754320 DOI: 10.1007/s00439-024-02718-6



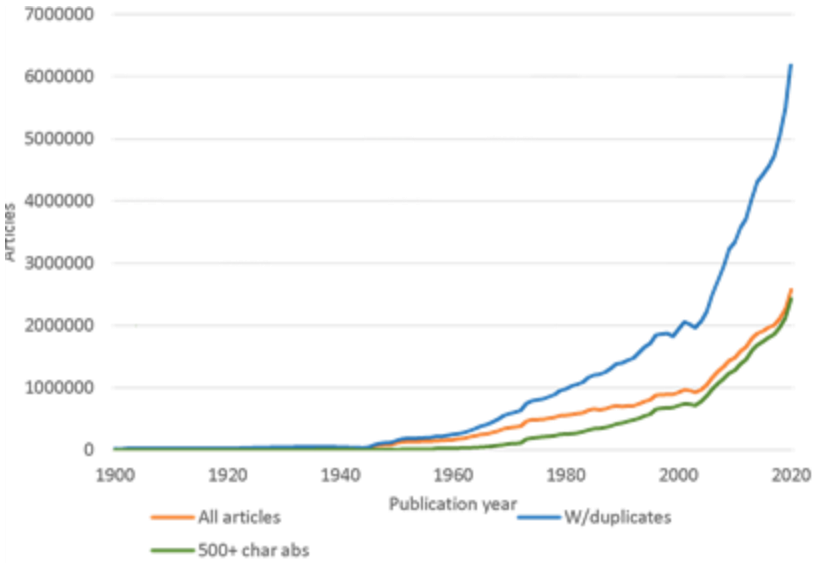
Panel v2

- NAV3
- ACO2
- GJC2
- POMK
- UPB1
- MYO5B

Scientific literature curation

Manual and time consuming

Literature corpus is growing continuously



adapted from *Scopus 1900–2020: Growth in articles, abstracts, countries, fields, and journals, Thelwall & Sud, 2022*

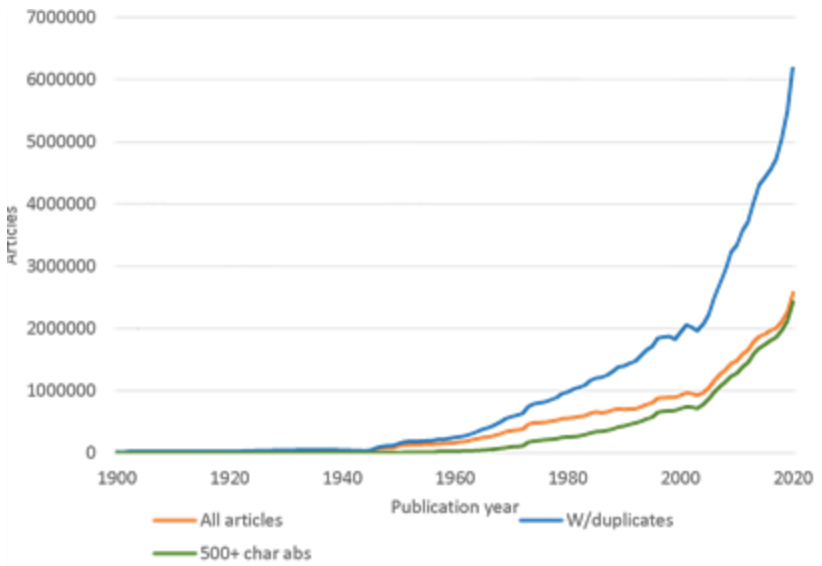
Covering a wide range of diseases:

Audiology
Cancer
Cardiology
Dermatology
Endocrinology
Gastrohepatology
Haematology
Immunology
Metabolic
Mitochondrial
Musculoskeletal
Neurology
Opthalmology
Renal
Respiratory

Scientific literature curation

Manual and time consuming

Literature corpus is growing continuously



adapted from *Scopus 1900–2020: Growth in articles, abstracts, countries, fields, and journals, Thelwall & Sud, 2022*

Covering a wide range of diseases:

Audiology
Cancer
Cardiology
Dermatology
Endocrinology
Gastrohepatology
Haematology
Immunology
Metabolic
Mitochondrial
Musculoskeletal
Neurology
Opthalmology
Renal
Respiratory

How can we search for literature in a better way?

Intellectual Disability panel

R29 Intellectual disability

Testing Criteria

Unexplained moderate/severe/profound global developmental delay or unexplained moderate/severe/profound intellectual disability, and where clinical features are suggestive of an underlying monogenic disorder requiring sequencing and targeted genetic testing is not possible.

A frequently applied panel for whole genome sequencing analysis

- Second **most applied** WGS panel in the NHS GMS
- Feeds into Paediatric disorders super panel – **most applied**

Challenges of this panel:

- Genetically **heterogenous** (> 1500 green genes already)
- **Frequent discoveries** of new genes and patients
- Often clinically **syndromic**
- **Diverse disease vocabulary** used in the literature

Hands-on task – how to find literature

Searching for literature

In groups at your tables discuss the following questions:

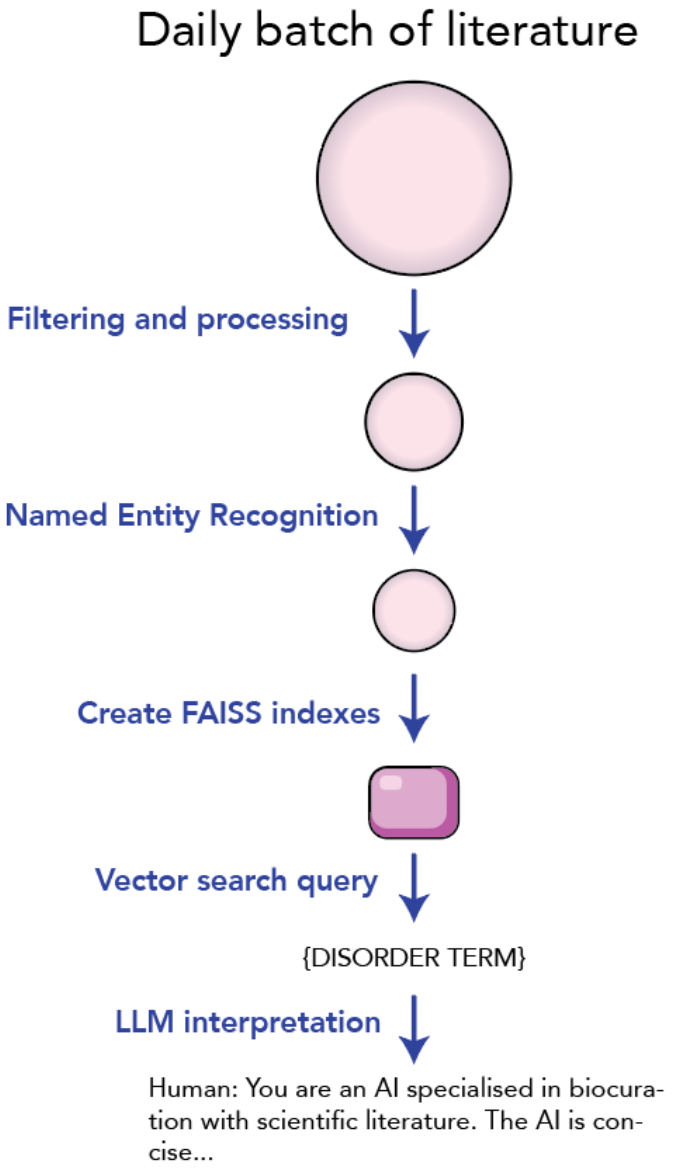
What tools do you use to find literature for your work?

What do you find difficult in finding literature?

Share your top answers

Machine learning and text processing

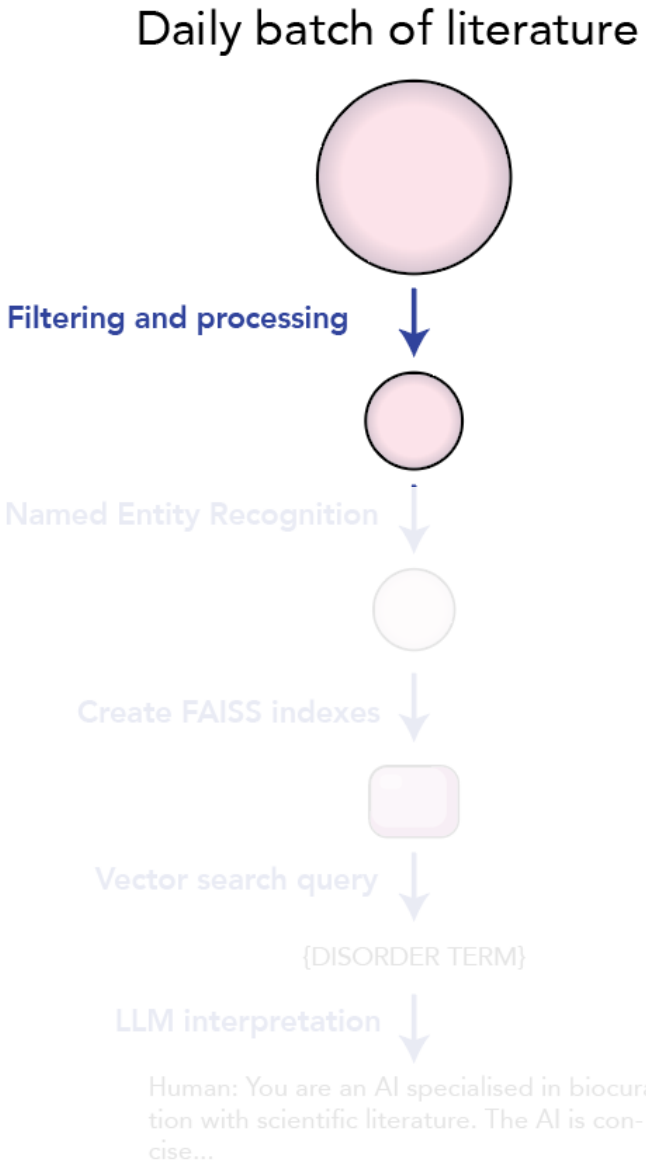
ML pipeline



Processing typically takes about 6 hours for a weeks worth (approx. 12,000) of papers

Approximately one positive paper per day found

ML pipeline



Remove non relevant journals: 884

e.g. Journal of Agricultural Food Chemistry, Frontiers in Veterinary Science

Remove certain publication types

e.g. books, reviews etc.

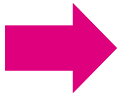
Removing bibliography

REFERENCES AND NOTES

1. T. Yue, H. Bloomfield-Gadêlha, J. Rossiter, Snail-inspired water-enhanced soft sliding suction for climbing robots. *Nat. Commun.* **15**, 4038 (2024).
   SEE ALL REFERENCES • CROSSREF • PUBMED • WEB OF SCIENCE • GOOGLE SCHOLAR
2. W. Pang, S. Xu, J. Wu, R. Bo, T. Jin, Y. Xiao, Z. Liu, F. Zhang, X. Cheng, K. Bai, H. Song, Z. Xue, L. Wen, Y. Zhang, A soft microrobot with highly deformable 3D actuators for climbing and transitioning complex surfaces. *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2215028119 (2022).
   CROSSREF • PUBMED • WEB OF SCIENCE • GOOGLE SCHOLAR
3. Y. Wu, X. Dong, J.-K. Kim, C. Wang, M. Sitti, Wireless soft millirobots for climbing three-dimensional surfaces in confined spaces. *Sci. Adv.* **8**, eab3431 (2022).
   GO TO REFERENCE • CROSSREF • WEB OF SCIENCE • GOOGLE SCHOLAR
4. C. Tang, B. Du, S. Jiang, Q. Shao, X. Dong, X.-J. Liu, H. Zhao, A pipeline inspection robot for navigating tubular environments in the sub-centimeter scale. *Sci. Robot.* **7**, eabm8597 (2022).
   SEE ALL REFERENCES • CROSSREF • PUBMED • WEB OF SCIENCE • GOOGLE SCHOLAR
5. B. Tao, Z. Gong, H. Ding, Climbing robots for manufacturing. *Natl. Sci. Rev.* **10**, nwad042 (2023).
   GO TO REFERENCE • CROSSREF • PUBMED • WEB OF SCIENCE • GOOGLE SCHOLAR

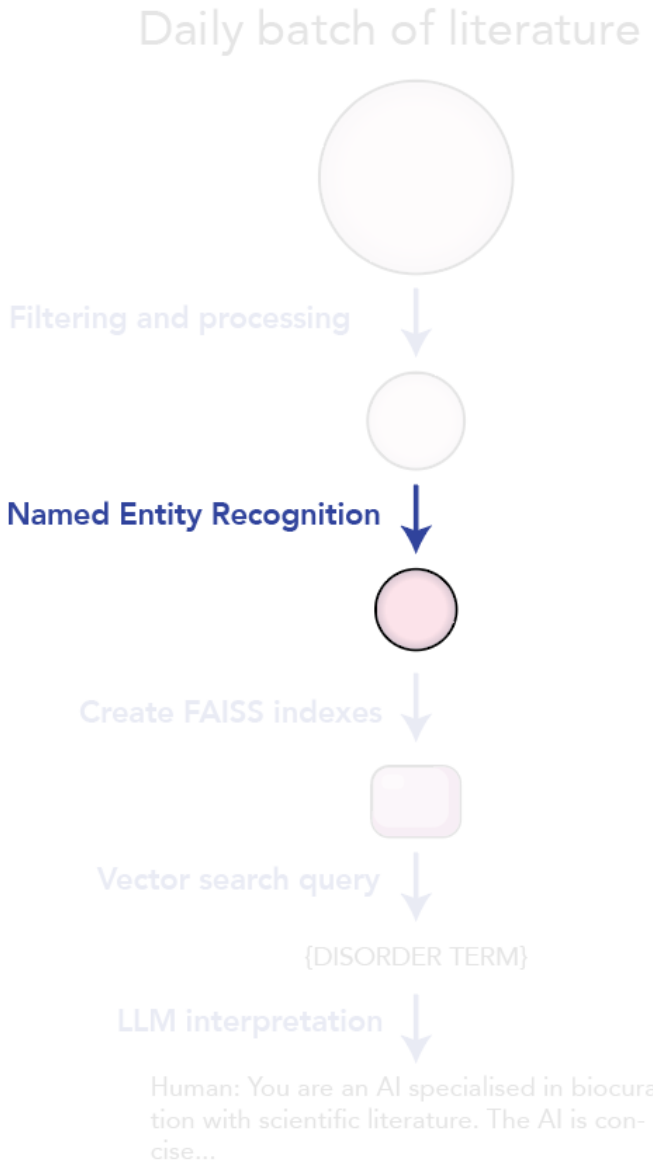
Cleaning up the data

<article xmlns:mml="http://www.w3.org/1998/Math/MathML" xmlns:xlink="http://www.w3.org/1999/xlink" article-type="research-article" dtd-version="1.1d3" xml:lang="en">\n<front>\n<journal-meta>\n<journal-id journal-id-type="nlm-ta">PLoS ONE</journal-id>\n<journal-id journal-id-type="publisher-id">plos</journal-id>\n<journal-id journal-id-type="pmc">plosone</journal-id>\n<journal-title-group>\n<journal-title>PLOS ONE</journal-title>\n</journal-title-group>\n<issn pub-type="epub">1932-6203</issn>\n<publisher>\n<publisher-name>Public Library of Science</publisher-name>\n<publisher-loc>San Francisco, CA USA</publisher-loc>\n</publisher>\n</journal-meta>\n<article-meta>\n<article-id pub-id-type="doi">10.1371/journal.pone.0222992</article-id>\n<article-id pub-id-type="publisher-id">PONE-D-18-34583</article-id>\n<article-categories>\n<subj-group subj-group-type="heading">\n<subject>Research Article</subject>\n</subj-group>\n<subj-group subj-group-type="Discipline-v3"><subject>People and places</subject><subj-group><subject>…



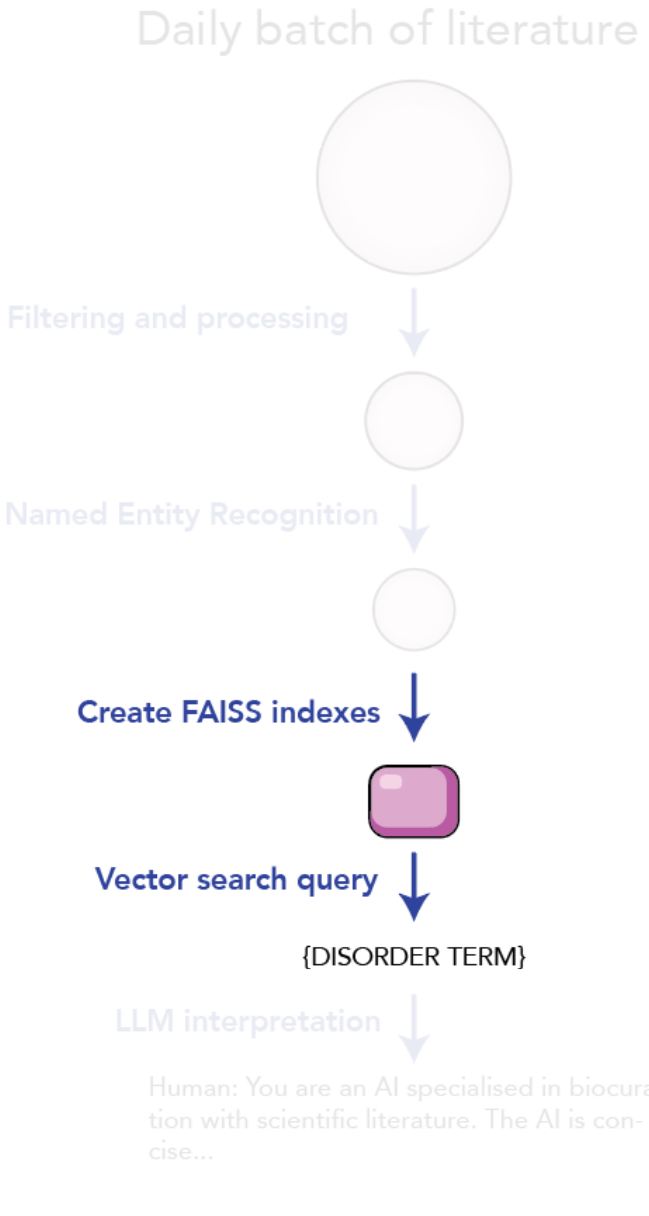
Background Impairments in social cognition have been described in several psychiatric and neurological disorders. Given the importance of the relationship between social cognition and functioning and quality of life in these disorders, there is a growing interest in social cognition remediation interventions. The aim of this study was to carry out a systematic mapping review to describe the state of the art in social cognition training and remediation interventions. Methods Publications from 2006 to 2016 on social cognition interventions were reviewed in four databases: Scopus, PsycINFO, PubMed and Embase. From the initial result set of 3229 publications, a final total of 241 publications were selected. Results The study revealed an increasing interest in social cognition remediation interventions, especially in the fields of psychiatry and psychology, with a gradual growth in the number of publications. These were frequently published in high impact factor journals and underpinned by robust scientific evidence. Most studies were conducted on schizophrenia, followed by autism spectrum disorders. Theory of mind and emotional processing were…

ML pipeline



Ceramides play a central role in human health and disease, yet their role as systemic signaling molecules remain poorly understood. In this work, we identify FPR2 fpr2 as a membrane receptor that specifically binds long-chain ceramides (C14-C20). In brown and beige adipocytes, C16:0 ceramide binding to FPR2 fpr2 inhibits thermogenesis via Gi gi -cyclic AMP signaling pathways, an effect that is reversed in the absence of FPR2 fpr2 . We present three cryo–electron microscopy structures of FPR2 fpr2 in complex with Gi gi trimers bound to C16:0, C18:0 and C20:0 ceramides. The hydrophobic tails are deeply embedded in the orthosteric ligand pocket, which has a limited amount of plasticity. Modification of the ceramide binding motif in closely related receptors, such as FPR1 fpr1 or FPR3 fpr3 , converts them from inactive to active ceramide ceramide receptors receptors . Our findings provide a structural basis for adipocyte thermogenesis mediated by FPR2 fpr2 .

ML pipeline



Convert text into machine readable embeddings

Allows searching of text via <u>semantic meaning</u>

<u>Intellectual disability</u>

Developmental disability
Cognitive impairment
Cognitive disability
Intellectual impairment
Neurodevelopmental disorder
Learning disability
General learning difficulty
Global developmental delay

Words mean different things in **context**

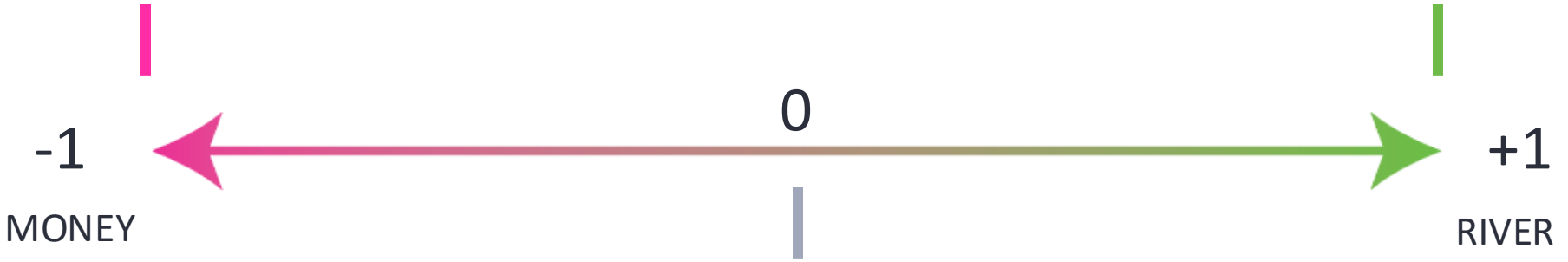
Sam went to the **bank** to get the money for his holiday

Sam enjoyed the view from the **bank** of the river

Machine learning representing a sentence

Sam went to the **bank** to get the money for his holiday

Sam enjoyed the view from the **bank** of the river

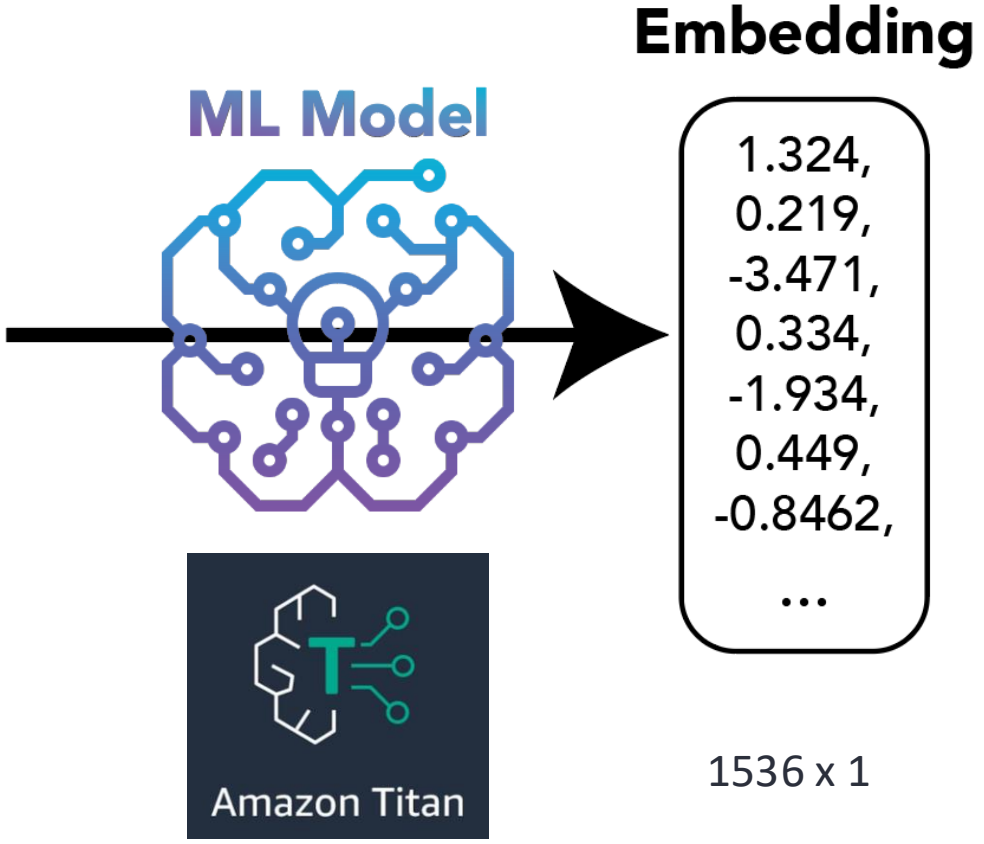


Sam gave a presentation

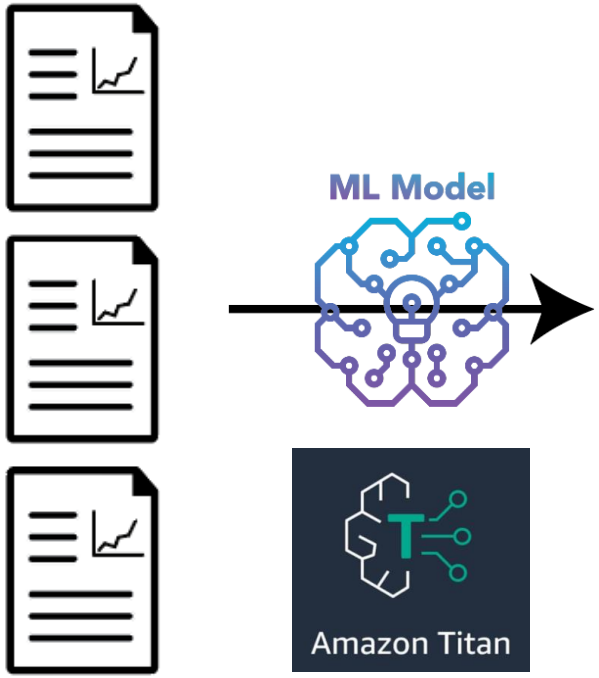
Representing the literature

Chunk of a paper

The acyl-CoA binding domain (ACB) and the ankyrin-repeat motifs (ANK) of ACBD6 can perform their functions independently. Interaction of ANK with human NMT2 was necessary and sufficient to provide protection. Fusion of the ANK module to the acyl-CoA binding protein ACBD1 was sufficient to confer the NMT-stimulatory property of ACBD6 to the chimera.



Representing the literature – Vectorstores



Vectorstore

1.324,	0.631,	0.394,	-3.125,	
0.219,	-2.111,	0.836,	-1.250,	
-3.471,	1.898,	2.659,	-0.365,	
0.334,	-0.319,	-3.145,	0.787,	...
-1.934,	0.005,	4.621,	-1.422,	
0.449,	-2.641,	-0.444,	1.179,	
-0.846,	-0.127	-0.213,	-0.987,	
...	...	...	...	

1536 x N

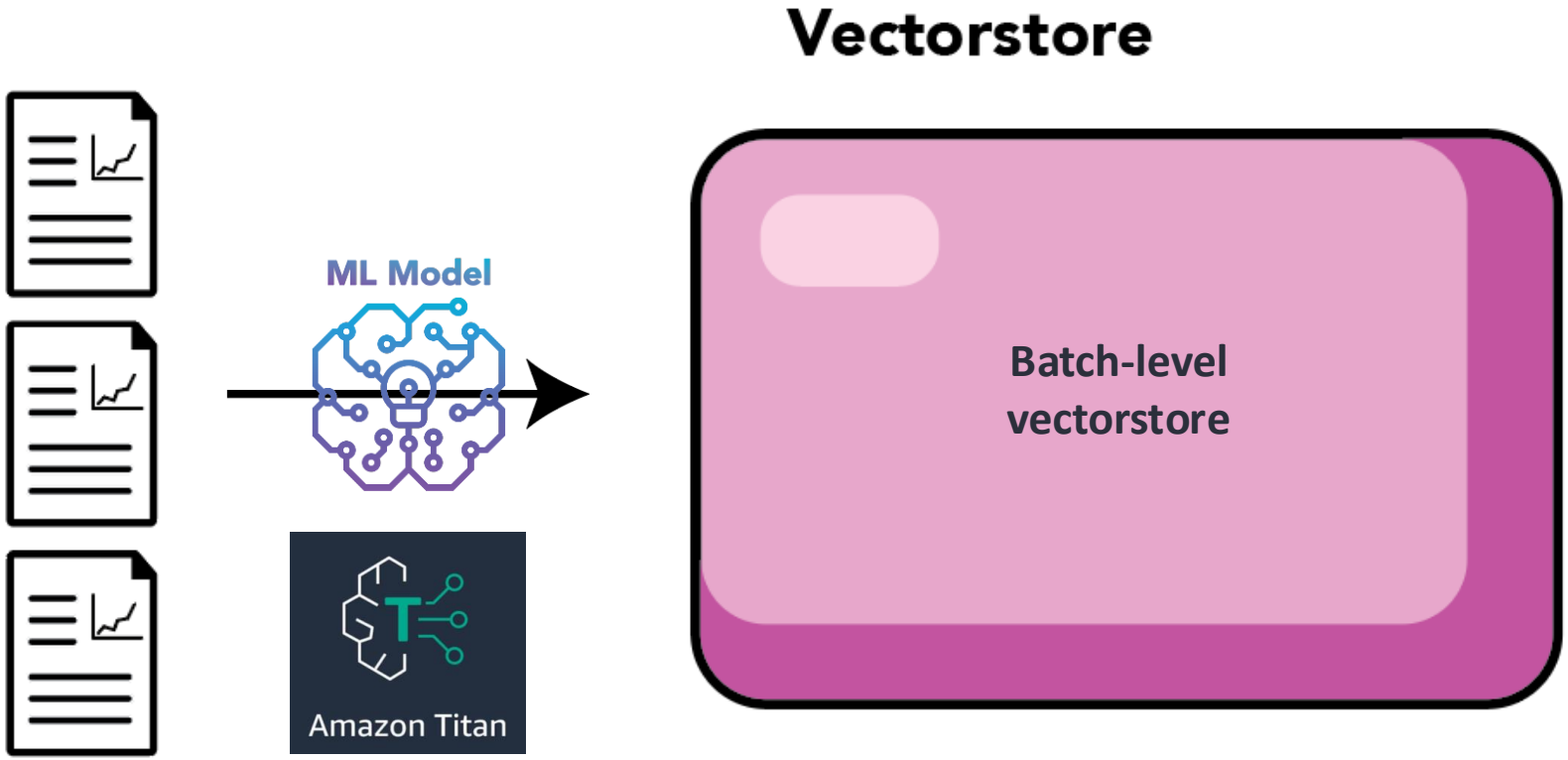
Multiple vectors per paper

Build one vectorstore per gene

Powered by FAISS*

**Facebook AI Similarity Search*

Representing the literature – Vectorstores



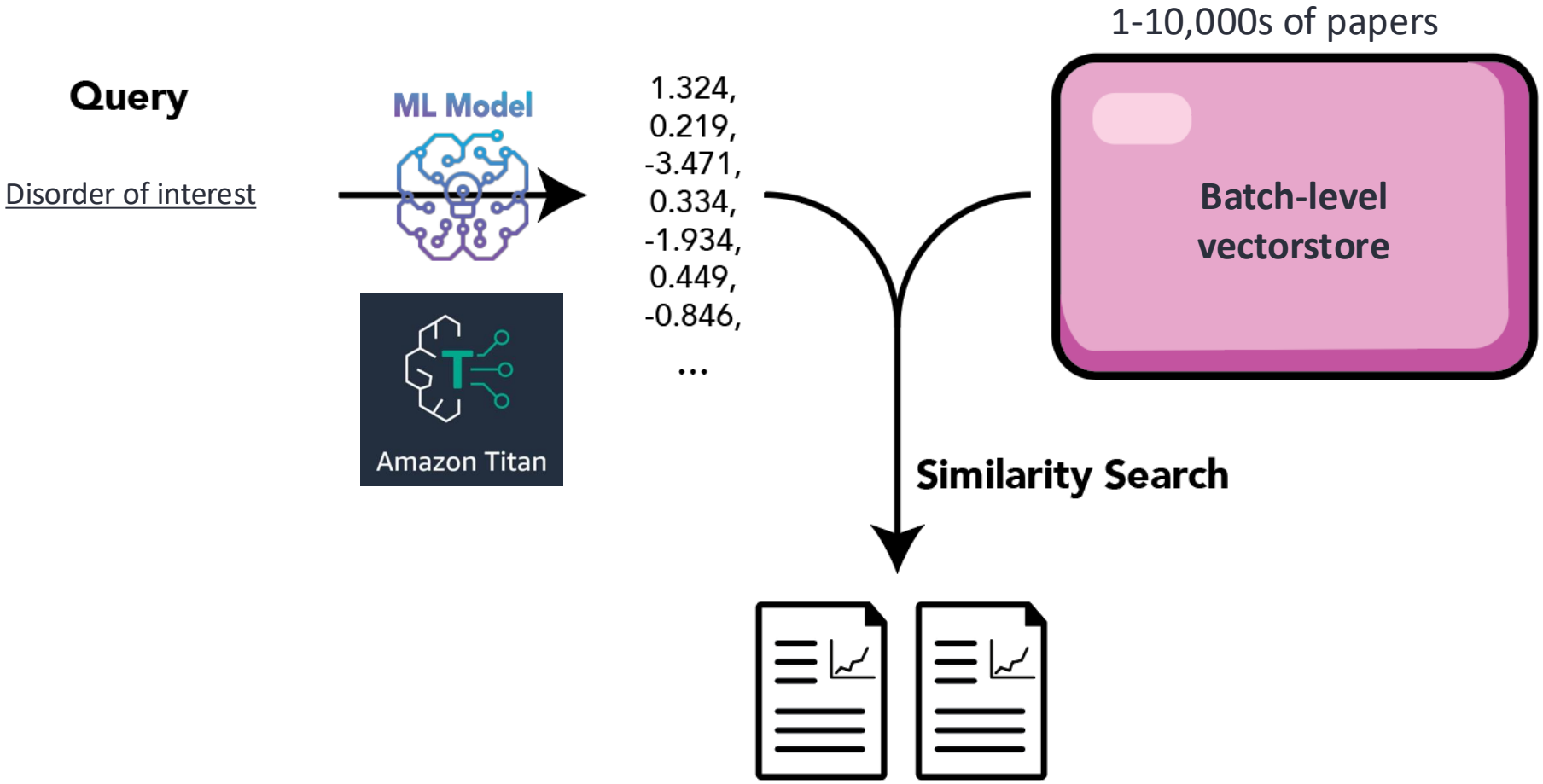
Multiple vectors per paper

Build one vectorstore per gene

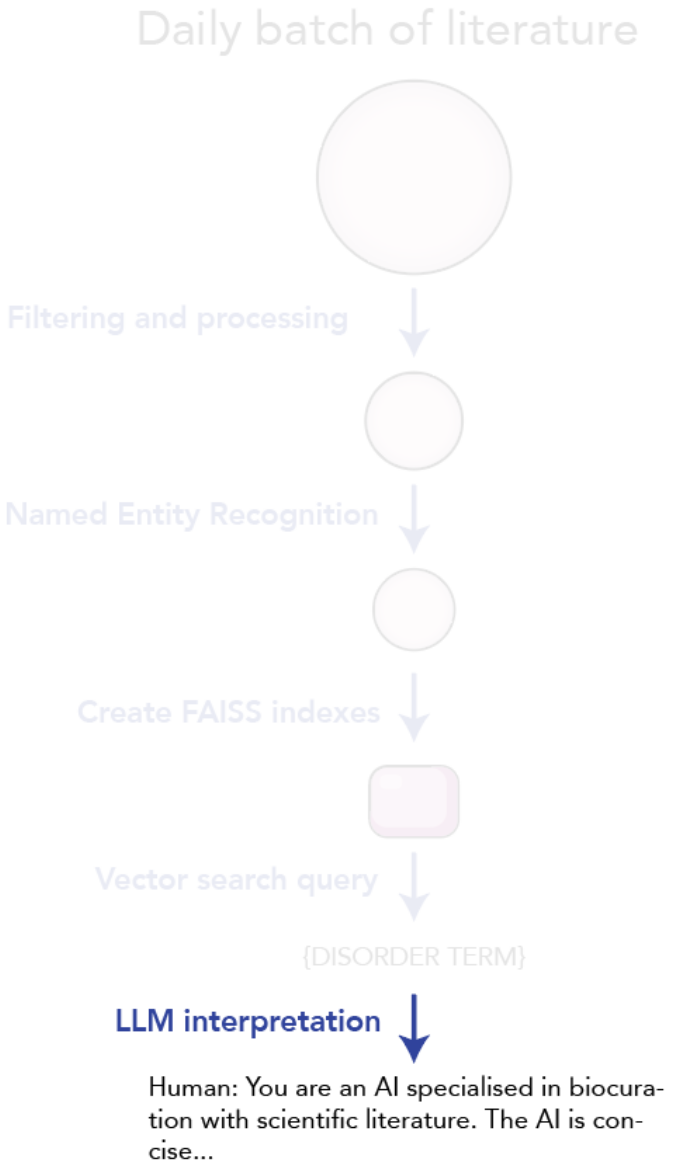
Powered by FAISS*

**Facebook AI Similarity Search*

Recovering potentially relevant literature



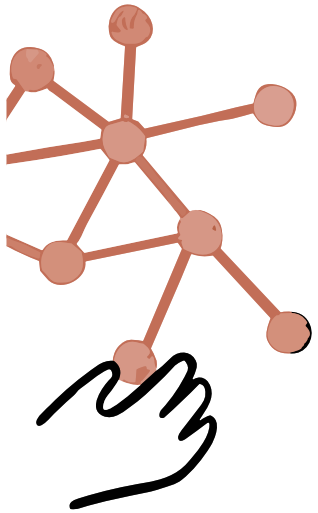
ML pipeline



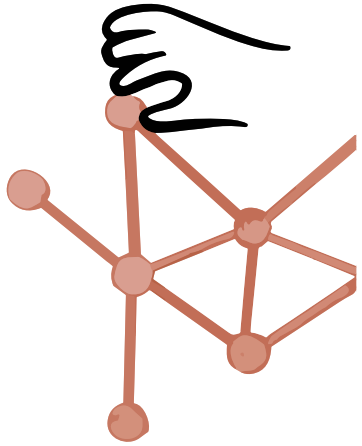
Pass entire paper to Large Language Model for interpretation

If model thinks paper is of value, extract the evidence and pass onto the Curation team

Utilising Large Language Model for analysis



Claude



ANTHROP\C

AWS Bedrock

No data protection issues since there is no PID involved in this data

Initial prompt

Human: You are an AI specialised in biocuration with scientific literature. The AI is concise and provides specific details from its context but limits it to 2000 tokens. If the AI does not know the answer to a question, it truthfully says it does not know.

Assistant: OK, got it, I'll be a biocuration specialised AI.

Human: Here is a paper from the scientific literature in <documents> tags:
<documents>
{document}
</documents>
Based on the above document, provide a short summary of the evidence, both direct and indirect even if it is inconclusive or not strong, linking mutation in {gene}, also known as {aliases}, to {disease} occurring in patients, probands or families. Answer only the words "don't know" if evidence is not present in the document. If the answer comes from a cited source, say that it is secondary ~~information and include the reference~~. Before answering, please think about the summary with <thinking></thinking> XML tags and place all other output in <answer> tags. ~~Remember~~ to only provide the summary if evidence occurs, otherwise the answer is only the words "don't know".

Assistant:<thinking>

Claude expects input to be in alternating "**Human**", "**Assistant**" format

Opening is often referred to as a **system prompt**

Insert a paper into the prompt

Claude was sometimes **overly cautious**

Don't rely on Claude implicitly knowing **gene aliases**

State whether source is **primary or secondary**

Allowing Claude to "**think**" improves performance

Output with **XML** allows **easy parsing**

Reaffirm to **only summarise** evidence, not the lack of

Kick off **response generation**

Updated prompt for structured output

Here is a paper from the scientific literature in <document> tags:
<document>
{document}
</document>
Based on the above document, extract the following information linking mutations to {disease} in if it is present and fill in the json schema below and place it between <json></json> tags, repeat the schema for each gene linked to {disease}.

schema:
<json>
{{"gene_name": //name of the gene\n, "segregation": // how does the gene segregate, choose from this list [”Full", ”Partial", ”Unknown"]\n, "sequencing_method": // choose from this list "WGS", "GWAS", "WES", "Panel", "Other", "Unknown"]\n, "phenotype_indicators": // be succinct in describing the patients phenotype\n, "source": /Is this a primary or secondary source of evidence, choose from ["Primary", "Secondary", "Unknown"]\n, "summary": //provide a text summary of relevant evidence from the paper}}
</json>

Only provide evidence occurring in human patients, probands or families. Answer only the words "I don't know" with no extra text if evidence is not present in the document. Before answering, please think about the summary with <thinking></thinking> XML tags and place all other output in <answer> tags.

Insert a paper into the prompt

Extract information on a **per-gene basis**

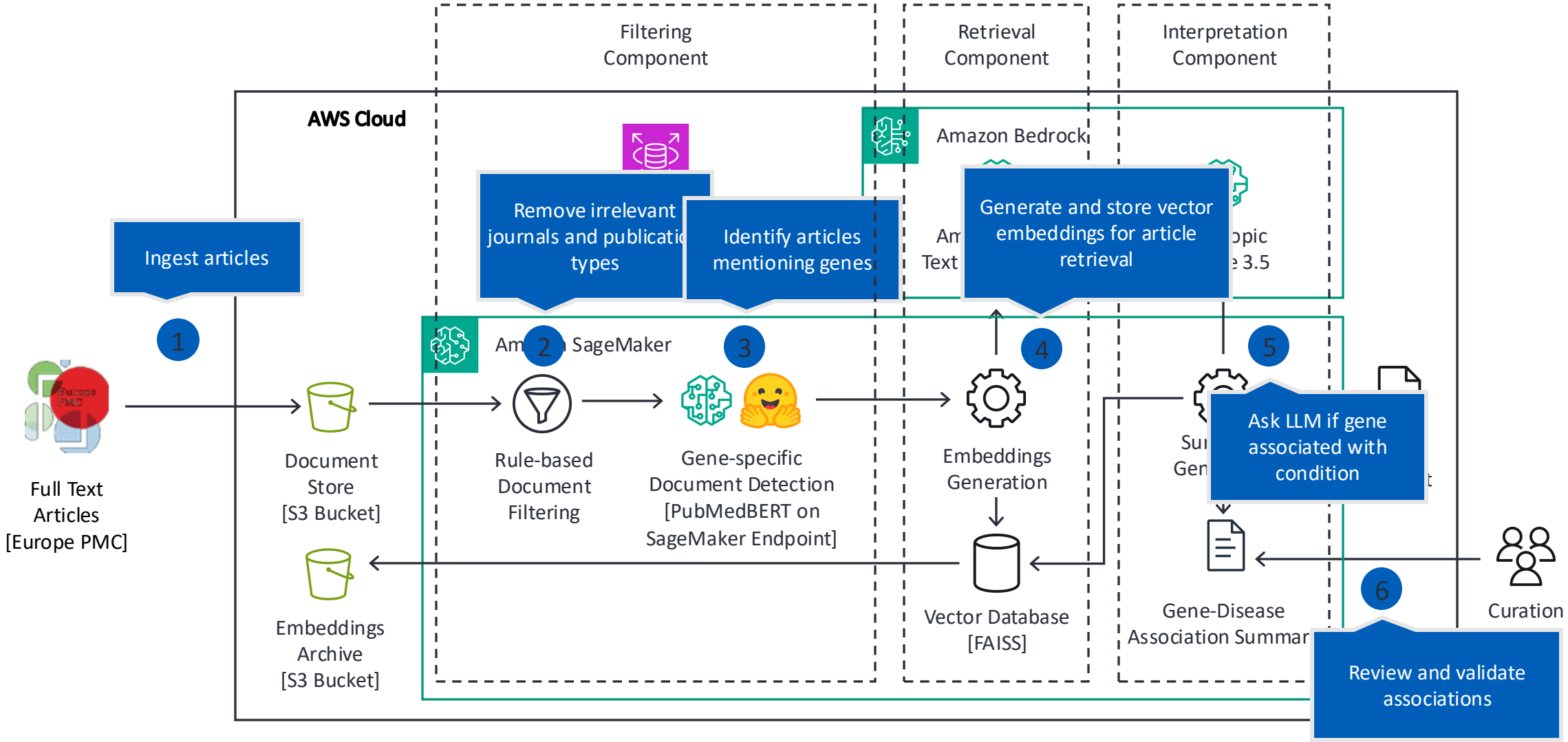
Extract in **JSON format**

Extract data using a **predefined list**

Reinforce Claude’s task and behaviour

Accelerating literature curation with GenAI on Amazon Bedrock

Approach & Solutions Architecture

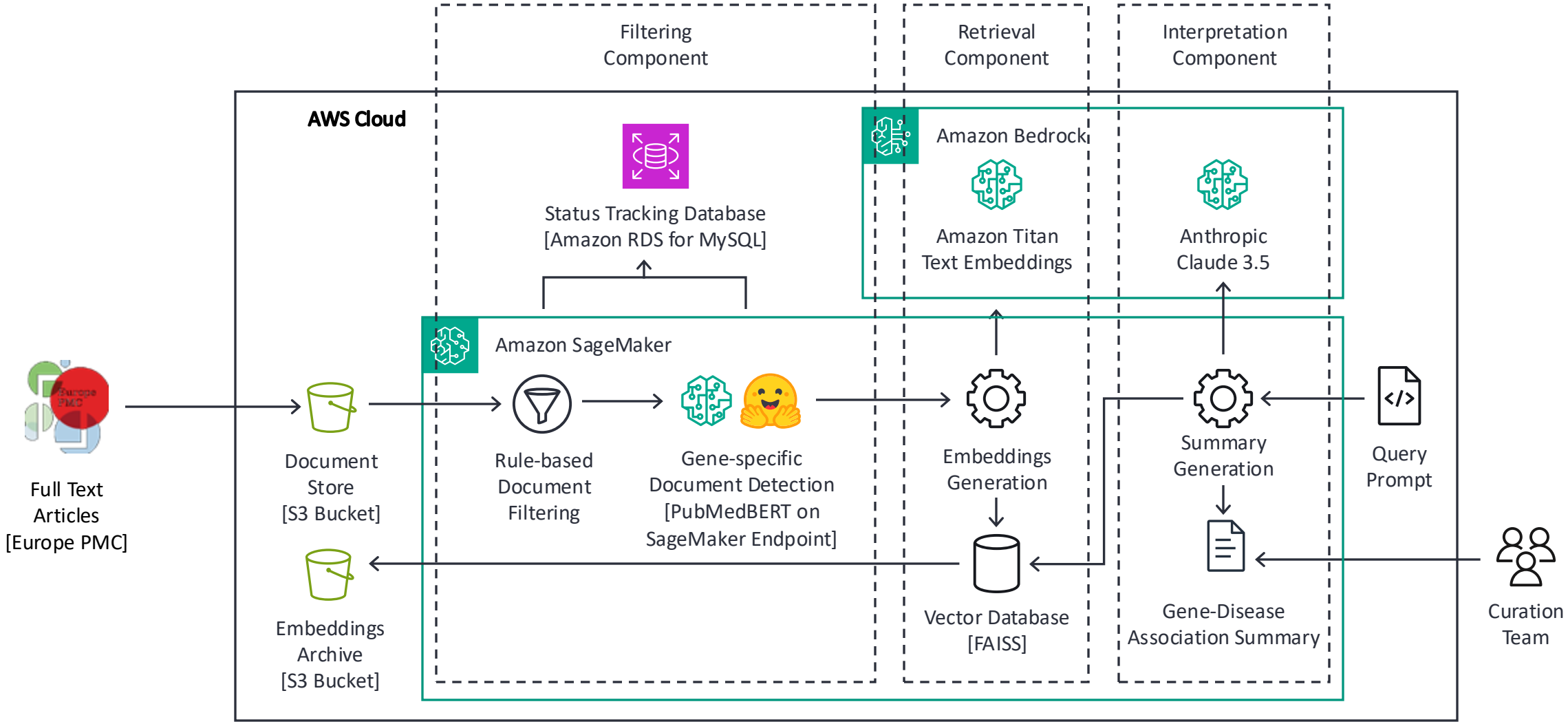


Genomics england

30

Accelerating literature curation with GenAI on Amazon Bedrock

Approach & Solutions Architecture



Genomics england

31

Demo

Questions?

Head-to-head challenge

Human vs machine

As a curator I want to find publications published in the last week that might provide evidence for gene-disease associations relating to Intellectual disability (moderate/severe/profound)

Your task: Search EuropePMC - https://europepmc.org/ - to find publications between the 1st and 10th February 2025 that might contain evidence for a gene-disease association.

("Intellectual disability" OR ”XX") AND (IN_EPMC:y) AND (FIRST_PDATE:[2025-02-01 TO 2025-02-10])

Make a table of relevant publications you find with:

- the PubMed/EuropePMC ID for the paper
- the gene linked to the intellectual disability phenotype
- the method used for identifying variants in the gene (e.g. WGS vs targeted panel)
- the number of cases and any segregation patterns reported within the families.
- the phenotypes of the patients
- a summary of the findings of the paper (stretch goal)

Results

How many publications did you find that looked useful?

Is there other information you think is useful for determining a gene-disease association?

What are the challenges in finding the information you wanted to record?

Results

How many publications did you find that looked useful?

Is there other information you think is useful for determining a gene-disease association?

What are the challenges in finding the information you wanted to record?

ML prioritized 12 gene-disease relationships from 11 publications

This took around 6 hours to run.

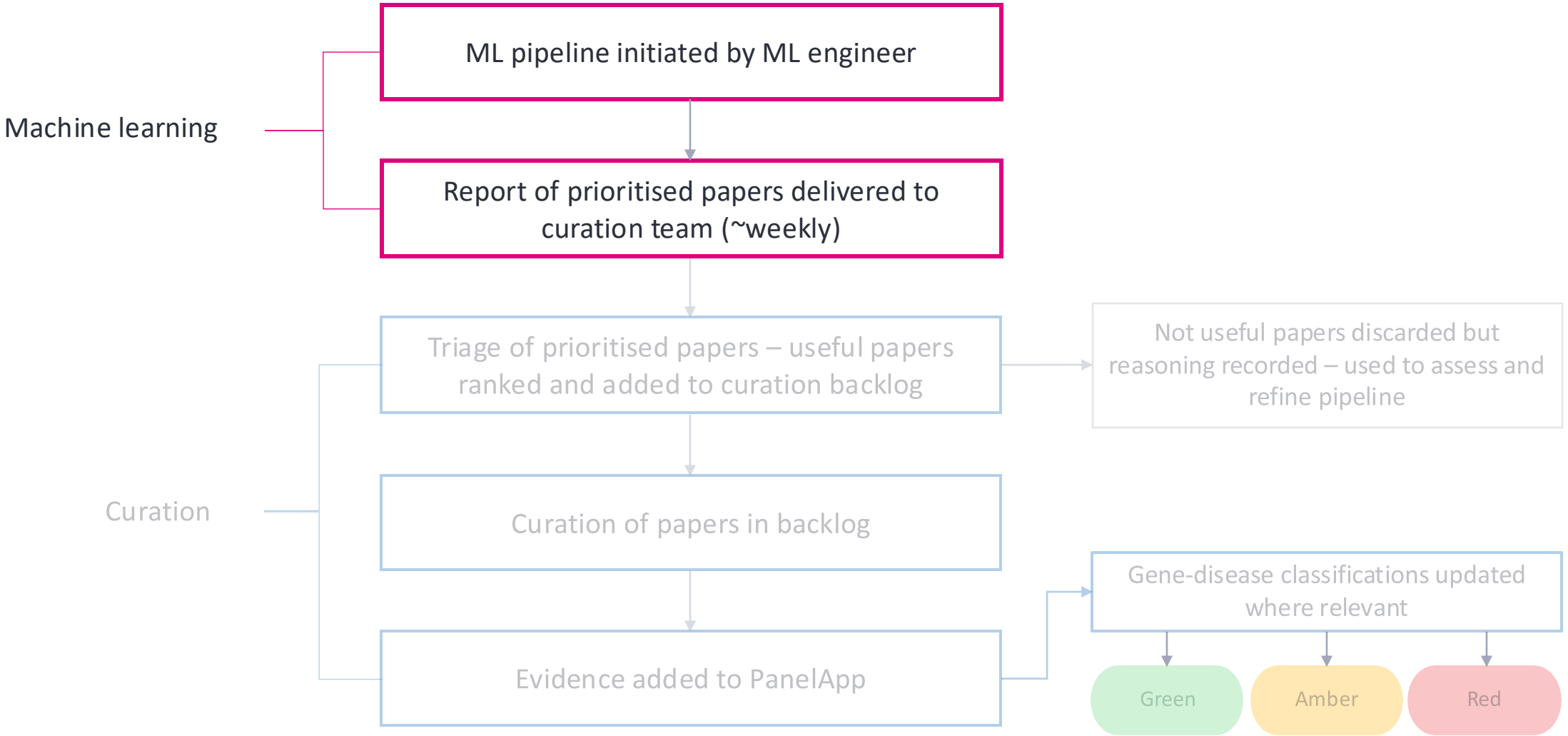
Gives a table of PMID, gene name, current rating of the gene on the panel, sequencing method, inheritance pattern, main phenotypes and a summary of the evidence.

From this:

- 5 genes were proposed to be promoted to green
- 1 gene promoted to amber
- 2 genes provided evidence for other panels (syndromic phenotype) but not ID

Was our ML experiment for searching for literature successful?

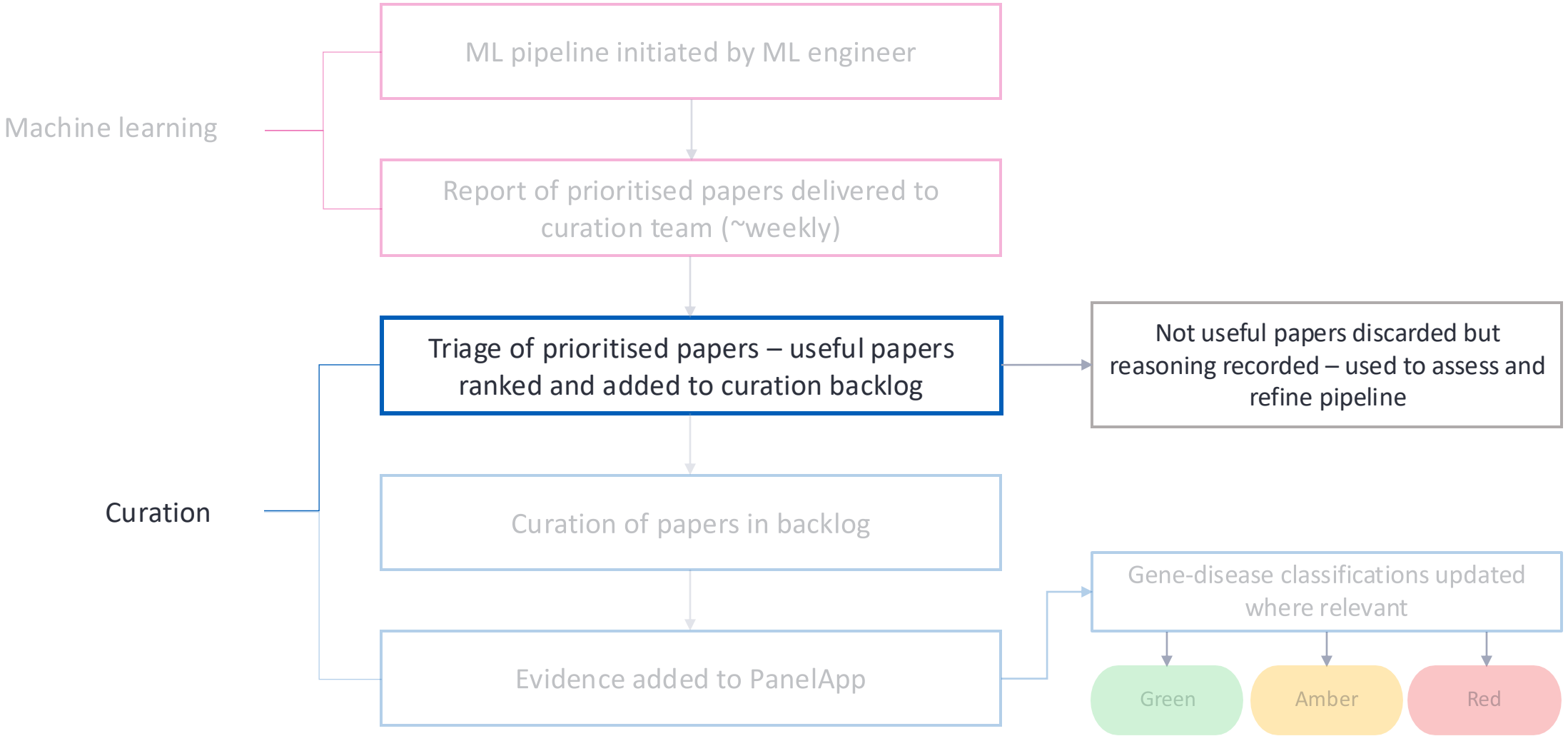
Overview of ML-Biocuration team workflow



Example Report

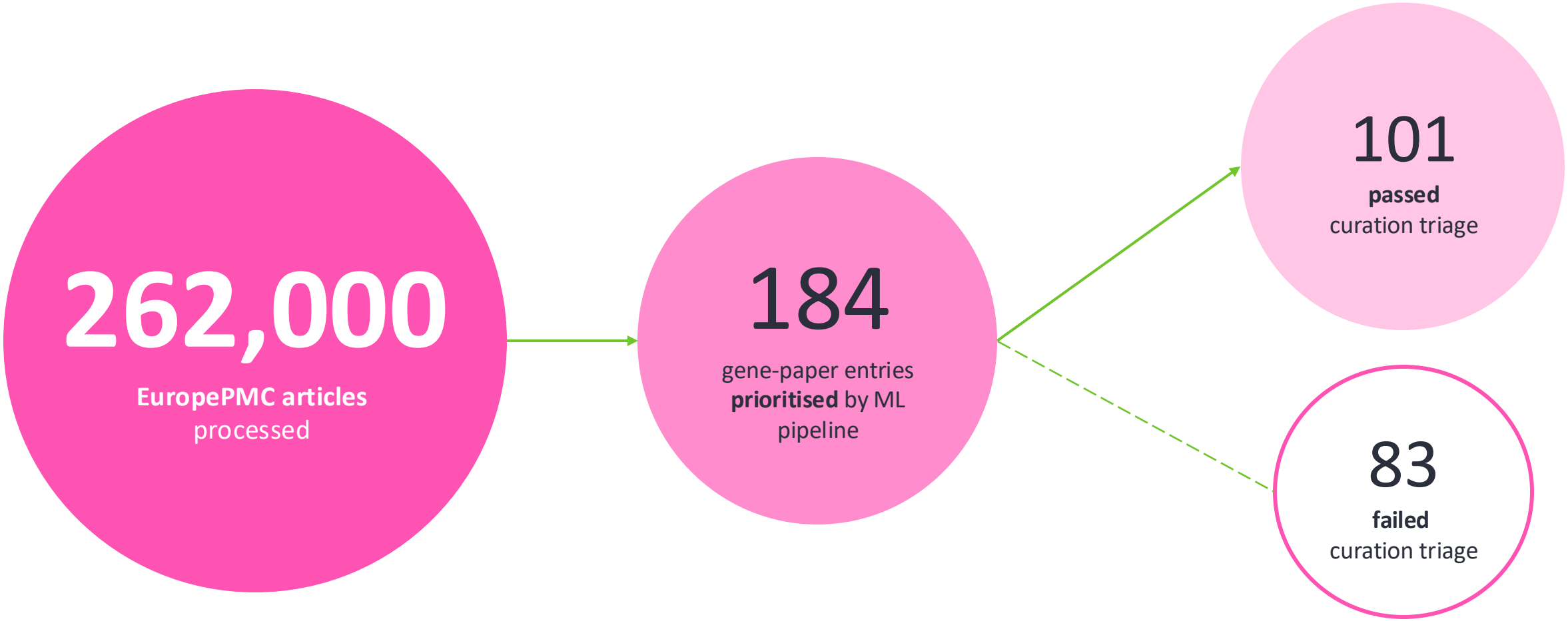
pmcid	pmid	gene_name	segregation	sequencing_method	phenotype_indicators	source	summary
PMC11341845	39174524	DENND5A	full	wes	Developmental delay, intellectual disability, drug-resistant seizures, microcephaly, ventriculomegaly, cerebral hypoplasia, hyperreflexia	primary	The paper describes a cohort of **24 individuals from 22 families with biallelic DENND5A variants who exhibit developmental and epileptic encephalopathy** (DEE), characterized by developmental delay, intellectual disability, and drug-resistant seizures. The individuals carry a range of DENND5A variants, including frameshift, nonsense, missense, intronic, and copy number variants. The **severity of intellectual disability and neurological abnormalities correlates with the type of DENND5A variant**, with biallelic frameshift or nonsense variants associated with more severe phenotypes. The paper provides evidence that biallelic loss-of-function mutations in DENND5A are causative for DEE and intellectual disability in this cohort.
PMC11343561	39184309	EIF2B3	unknown	wes	Global developmental delay	primary	The study identified a **missense variant (c.1103C>T, p.S368L) in the EIF2B3 gene in a patient with global developmental delay**, suggesting a potential link between this variant and intellectual disability.
PMC11341004	39176129	PTRH2	full	wes	moderate intellectual disability, motor development delay, hearing loss, peripheral neuropathy, ataxia, foot and facial dysmorphic features, pancreatic insufficiency	primary	The paper reports **two sisters of Iranian origin with a homozygous missense likely pathogenic variant c.254A>G, p.(Gln85Arg) in the PTRH2 gene**, confirmed by whole-exome sequencing and Sanger sequencing. Both sisters presented with **moderate intellectual disability** along with other symptoms characteristic of infantile-onset multisystem neurologic, endocrine, and pancreatic disease type 1 (IMNEPD1) caused by biallelic PTRH2 variants.
PMC11340112	39169373	IARS2	full	wes	Leigh syndrome, developmental delays, seizures, brain MRI abnormalities	primary	The study identified **compound heterozygous mutations in the IARS2 gene in two unrelated patients with Leigh syndrome** through whole exome sequencing. **Functional studies** showed that these mutations led to decreased IARS2 protein levels and impaired mitochondrial function due to deficiencies in OXPHOS complexes I and III, providing evidence for the pathogenicity of the identified IARS2 mutations in causing Leigh syndrome.

Overview of ML-Biocuration team workflow



ML identified papers

In 6 months, focusing on the Intellectual Disability panel...



Reasons for discarding a paper

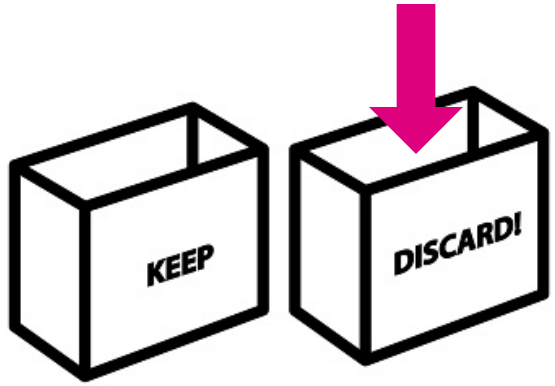
- **Disorder did not align with panel scope**
  - e.g. mild ID, neurodegeneration, ID caused by seizures
  - The curator, rather than the model, should make this judgment

- **Patient carried alterations in other genes**
  - Implication of any singular gene was unclear
  - The curator, rather than the model, should make this judgment

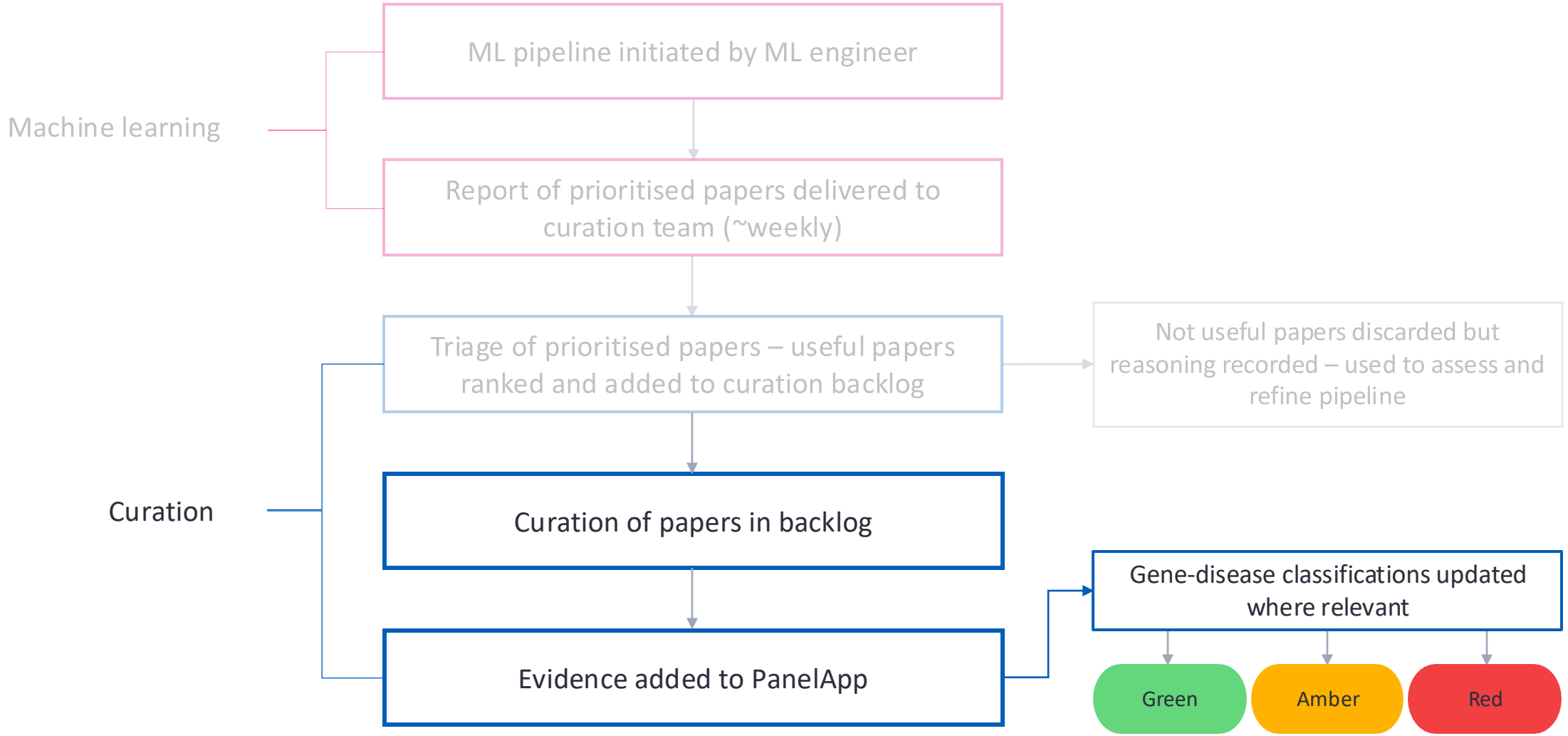
- **Secondary source, referencing another, already curated paper**
  - Rarely useful, so now exclude these

- **Gene already on panel**
  - Gene aliases, locus, protein, complex names
  - Potential for future refinement

- **Paper not relevant to discovery of rare gene-diseases**
  - GWAS, literature review, poor quality publication
  - Potential for future refinement

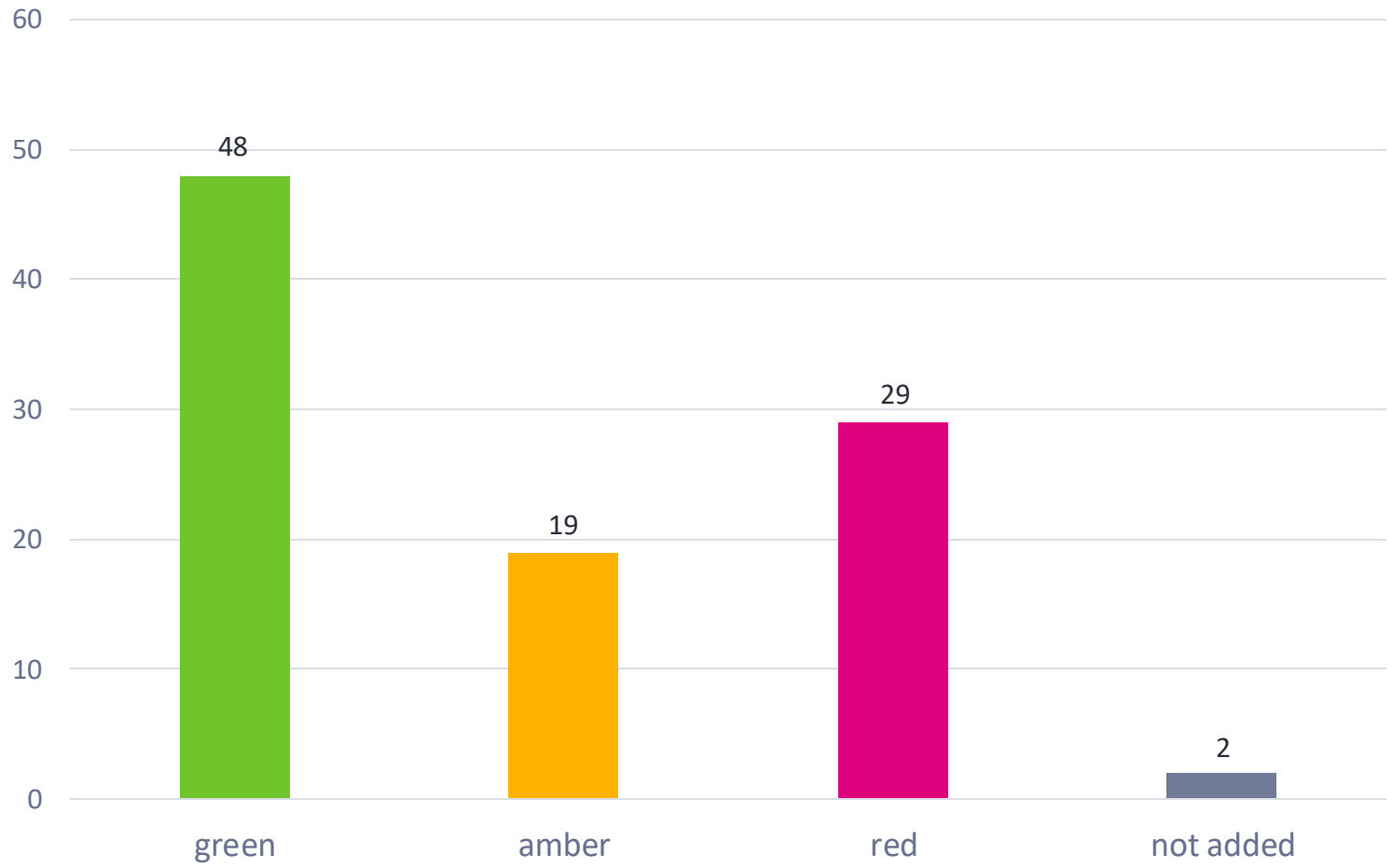


Overview of ML-Biocuration team workflow



Curation Results

101
passed
curation triage

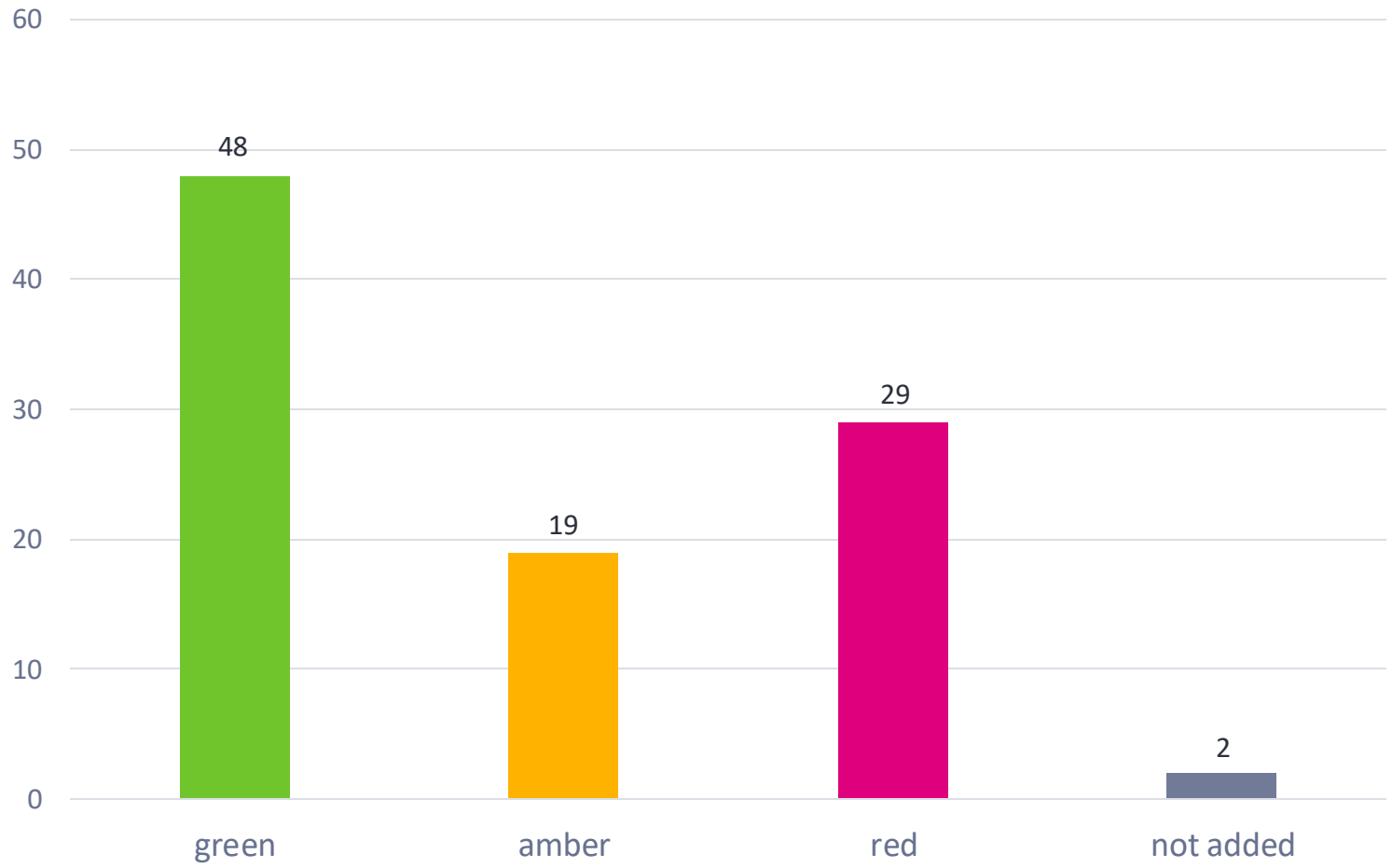


Highest rating based on ML identified publication

Total =
98 unique genes

Curation Results

101
passed
curation triage



Highest rating based on ML identified publication

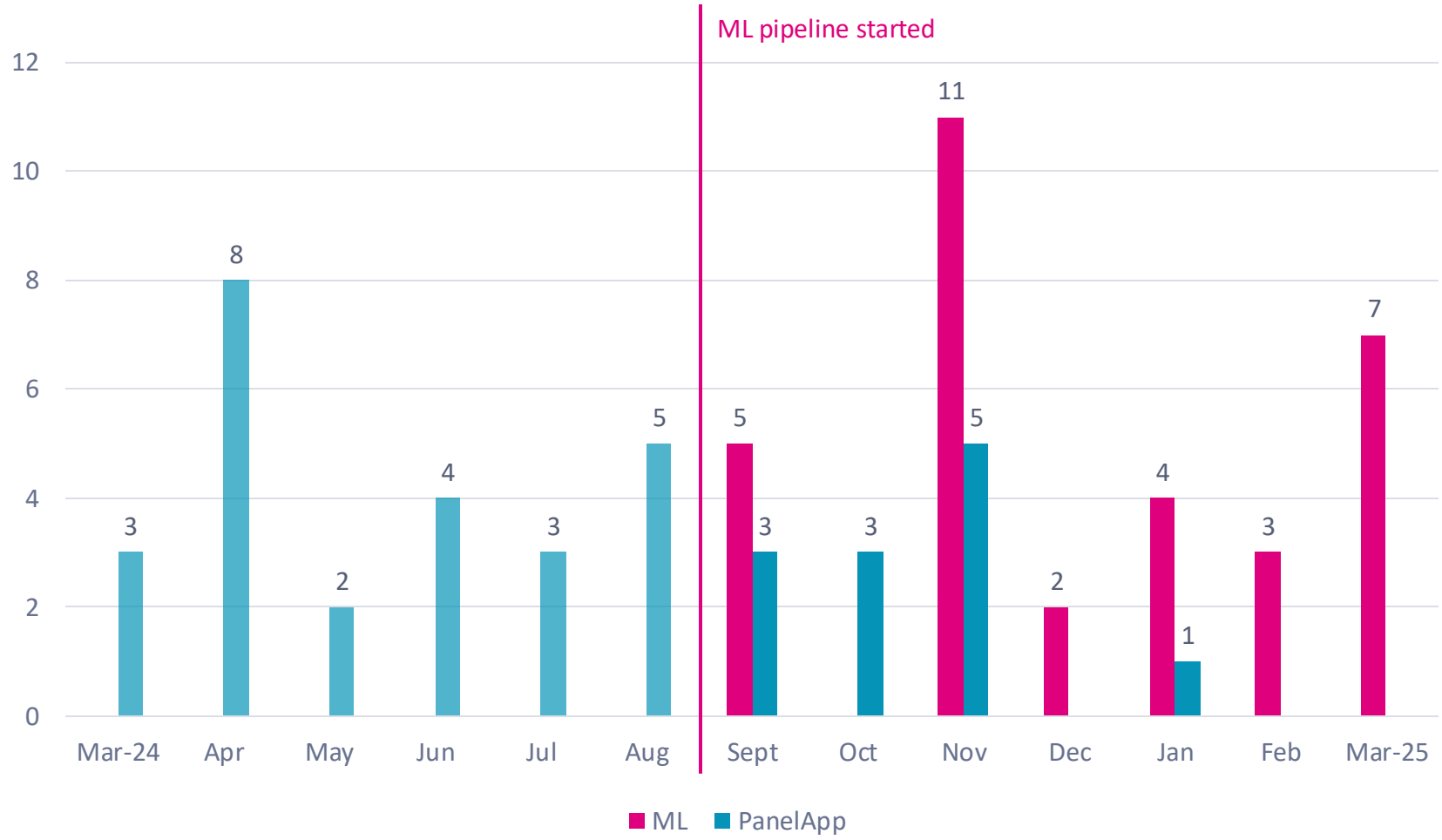
Found diagnostic level of evidence for genes – what we aimed to do

Total =
98 unique genes

ML vs crowdsourcing evidence

Sources leading to addition of a new green gene on ID panel

Monthly average:
- PanelApp: 2.8 genes
- ML: 4.6 genes



3 genes were identified by both methods

PanelApp = expert reviews submitted to PanelApp (including NHS) and Diagnostic Discovery

Key takeaways

Text mining is a valuable tool for biocuration and can increase productivity

There are limitations due to literature behind paywall, biases towards making associations, and time cost setting up and refining pipeline

Future directions

Explore other use cases:

- Search for new literature for low evidence genes on the panel
- Expand to other disease areas
- Proposing panels a gene can be added to

ML is a powerful tool, but not a replacement for human expertise

Questions?

A Genomics England – AWS partnership

Special thanks to our team

Genomics England

Francisco Azuaje
Arina Puzriakova
Achchuthan Shanmugasundram
Catherine Snow
Applied ML and Biocuration teams

AWS

Lou Warnett
Cemre Zor
Michael Mueller
Pablo Nuñez Pölcher
Dave Warke
Matt Howard