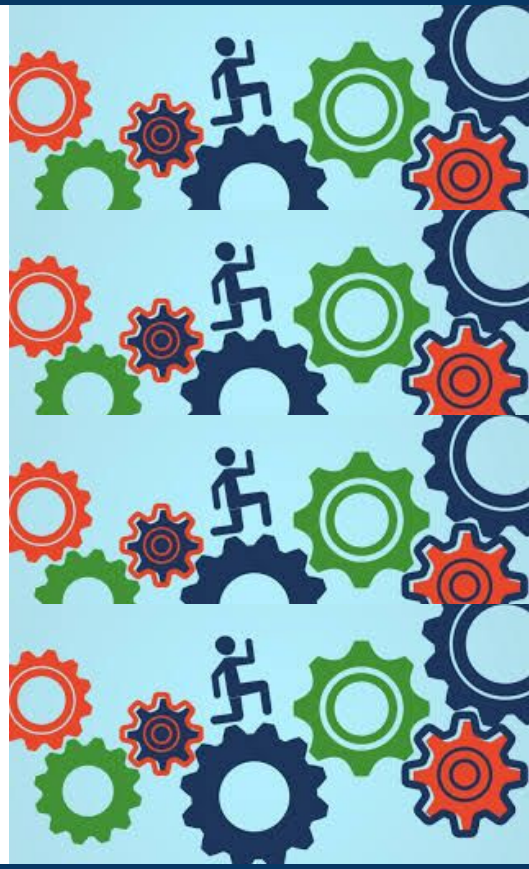
Extending the Gene Ontology for Biological Network Modeling

Pascale Gaudet
SIB Swiss Institute of Bioinformatics
GO Central

Gene Ontology Knowledgebase: Contents and optimization



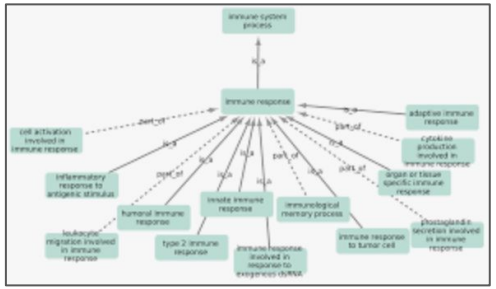
The Gene Ontology (GO)

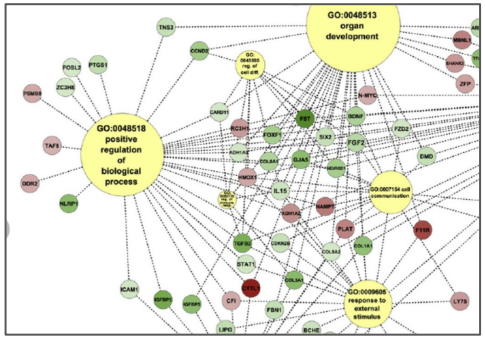
Provides a computational model of **gene function** at every level:

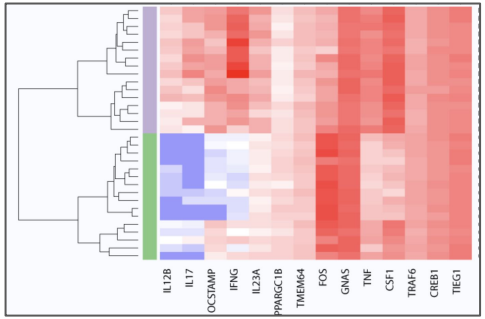
- molecular activities
- signaling and metabolic pathways
- cellular organism-level systems

Consists of:

1) **ontology**: a framework to describe the biological concepts
2) **annotations**: associations between genes and GO terms







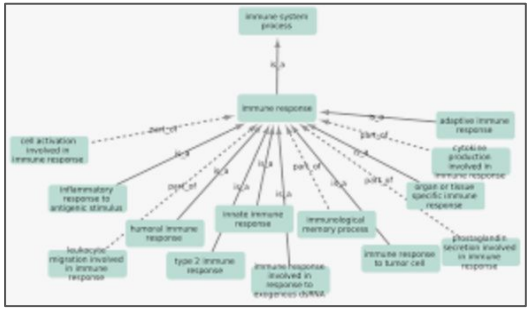
Current GO knowledgebase content

January 2026 release

Ontology

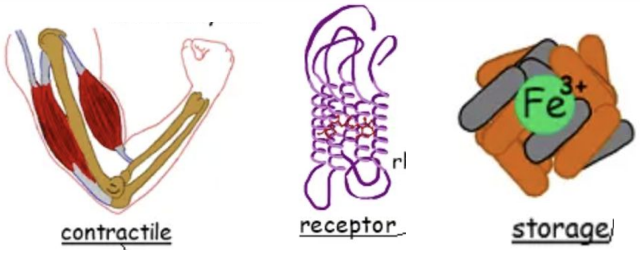
38,739 terms

- 24,547 BP
- 10,123 MF
- 4,069 CC



Annotations

- > 1 M annotations based on experimental data
- Millions of annotations by phylogenetic and electronic methods (800M in QuickGO/UniProt)



The 3 aspects of GO

Molecular Function (MF)

- ***Biochemical activity*** that a gene product performs
- Examples: enzymatic activity, adaptor activity, transcription factor activity

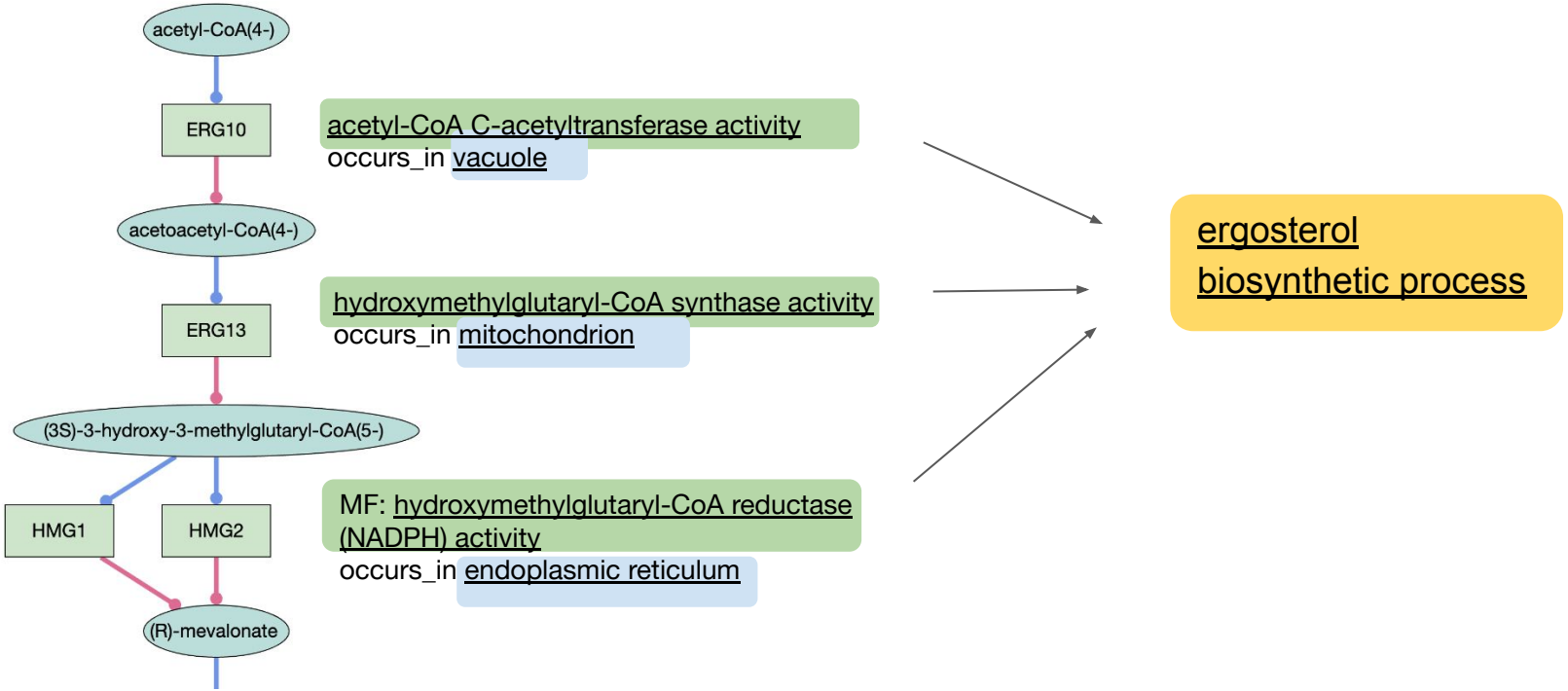
Cellular Component (CC)

- Cellular location where a gene product performs its activity (MF)
- Example: nucleus, cytoskeleton, ribosome

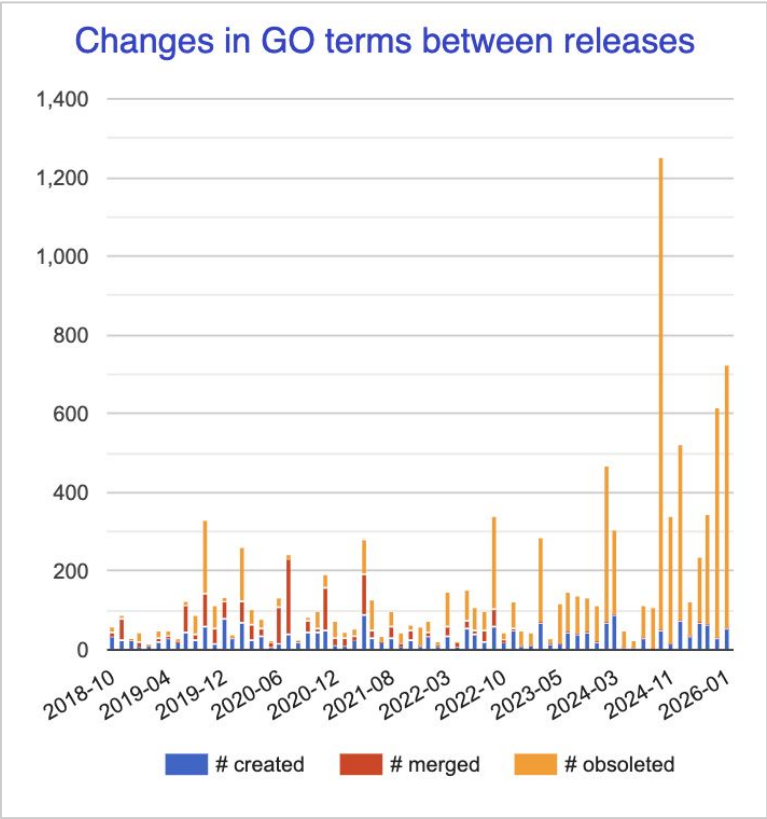
Biological Process (BP)

- Series of events making up a ***biological module or program***, accomplished by a specific set of molecular functions, in a specific order
- Examples: glycolysis, transcription, photosynthesis

Example: function, component and process



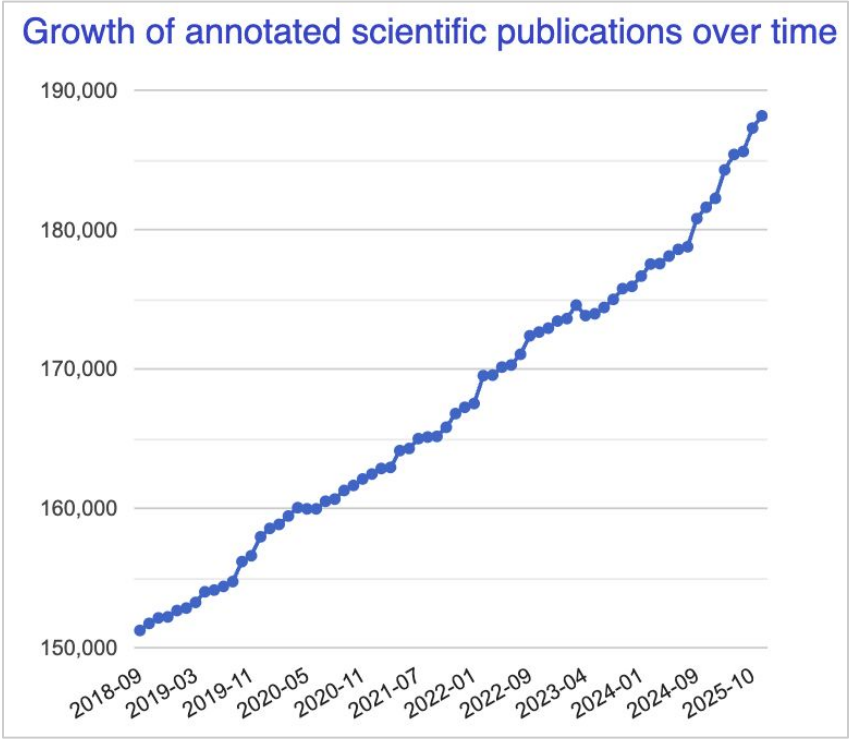
GO is dynamic



- GO aims to represent the current state of knowledge in biology
- Constantly revised as knowledge evolves
- Most change requests come from biocurators when annotating papers
- Large refactorings are made collaboratively with domain experts in particular areas of biology

Annotations are added and removed to represent current state of knowledge

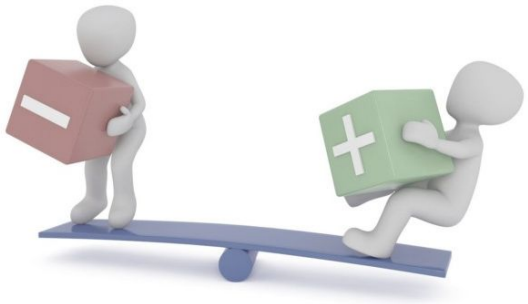
- A GO annotation is a statement that links a gene product and a GO term
- These are linked via relations from the Relations Ontology
- GO aims to represent the **current state of knowledge** in biology, hence it is constantly revised and expanded as biological knowledge accumulates



GO annotations

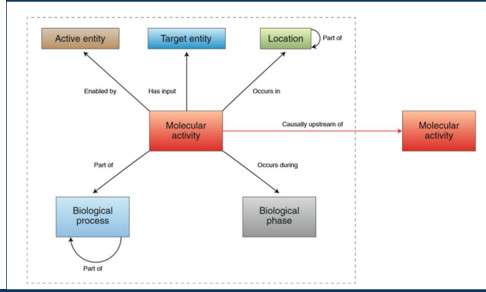
- Each gene can have any number of annotations to represent:
  - its function(s) and substrates
  - the cellular locations where it is active
  - the roles it plays in any process
  - each paper and evidence

- Both a feature and a limitation



Representing the complexity of biology:
from lists to models

Gene Product	Symbol	Qualifier	GO Term	Evidence	Reference
UniProtKB:P42697	DRP1A	acts_upstream_of_or_within	GO:0000911 cytokinesis by cell plate formation	ECO:0000316 IGI	PMID:12671086
UniProtKB:P42697	DRP1A	acts_upstream_of_or_within	GO:0009793 embryo development ending in seed dormancy	ECO:0000316 IGI	PMID:12671086
UniProtKB:P42697	DRP1A	acts_upstream_of_or_within	GO:0009920 cell plate formation involved in plant-type cell wall biogenesis	ECO:0000316 IGI	PMID:12671086
UniProtKB:P42697	DRP1A	acts_upstream_of_or_within	GO:0010051 xylem and phloem pattern formation	ECO:0000315 IMP	PMID:15923323
UniProtKB:P42697	DRP1A	acts_upstream_of_or_within	GO:0010091 trichome branching	ECO:0000315 IMP	PMID:12671086
UniProtKB:P42697	DRP1A	acts_upstream_of_or_within	GO:0048766 root hair initiation	ECO:0000315 IMP	PMID:27251533



From function to GO annotation

P46934 · NEDD4_HUMAN

Protein[i]	E3 ubiquitin-protein ligase NEDD4
Gene[i]	NEDD4

➢ **Molecular activity**

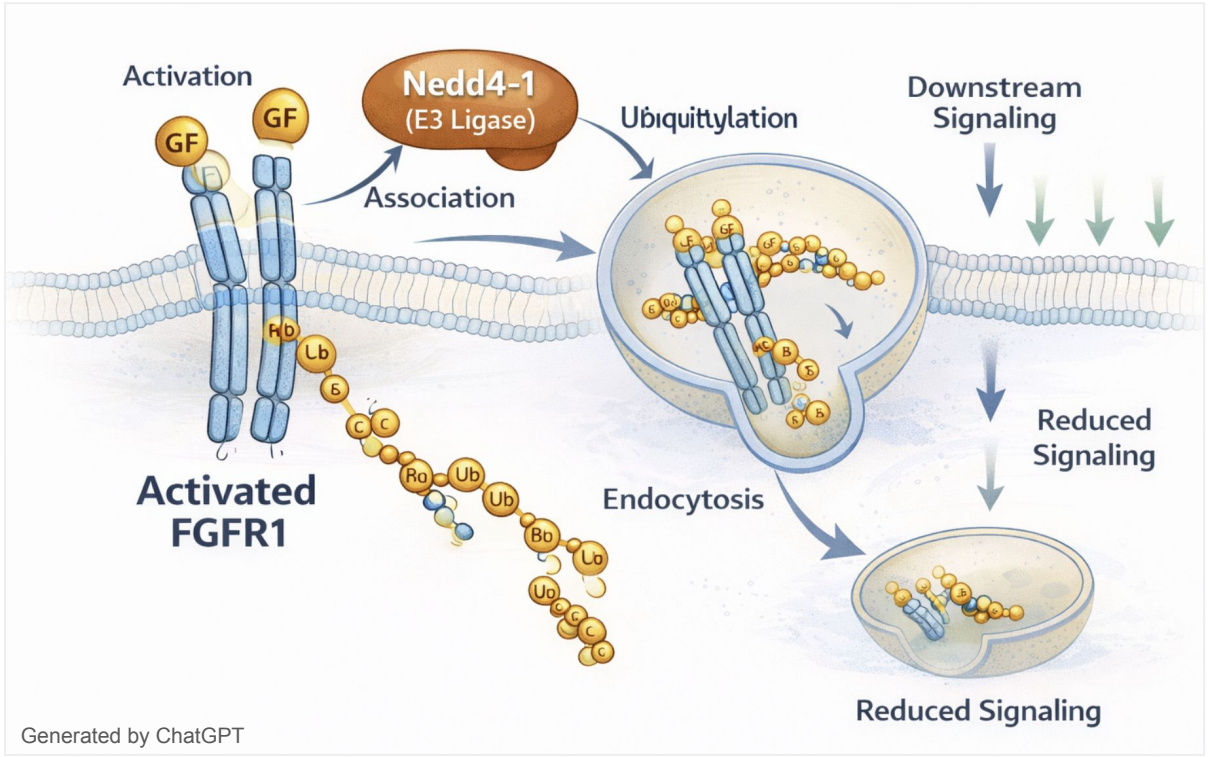
Function[i]

E3 ubiquitin-protein ligase which accepts ubiquitin from an E2 ubiquitin-conjugating enzyme in the form of a thioester and then directly transfers the ubiquitin to targeted substrates. Specifically ubiquitinates 'Lys-63' in target proteins (PubMed:19920177, PubMed:21399620, PubMed:23644597).
Involved in the pathway leading to the degradation of VEGFR-2/KDFR, independently of its ubiquitin-ligase activity. Monoubiquitinates IGF1R at multiple sites, thus leading to receptor internalization and degradation in lysosomes (By similarity).

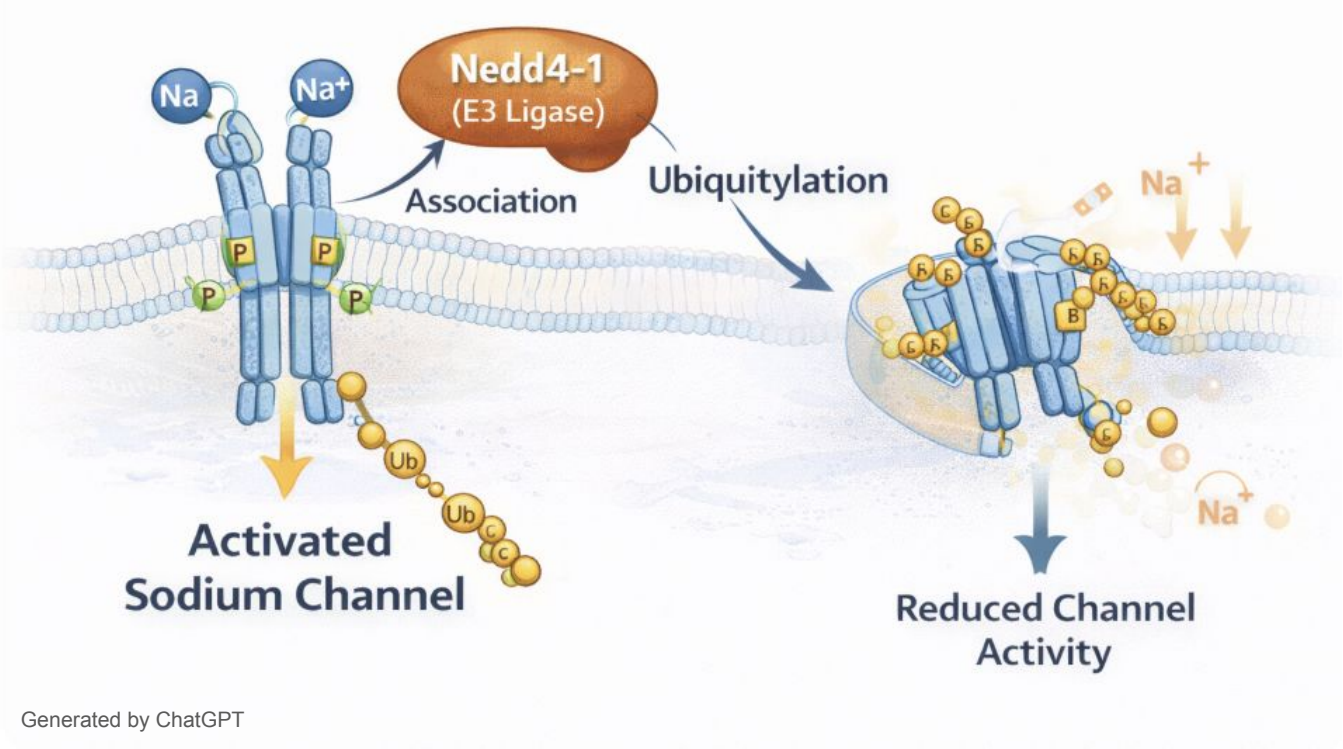
Many substrates with roles in various cellular processes

Ubiquitinates TNK2 and regulates EGF-induced degradation of EGFR and TNF2 (PubMed:20086093).
Ubiquitinates BRAT1 and this ubiquitination is enhanced in the presence of NDFIP1 (PubMed:25631046).
Ubiquitinates DAZAP2, leading to its proteasomal degradation (PubMed:11342538).
Ubiquitinates POLR2A (PubMed:19920177).

NEDD4 downregulates FGFR1 signaling

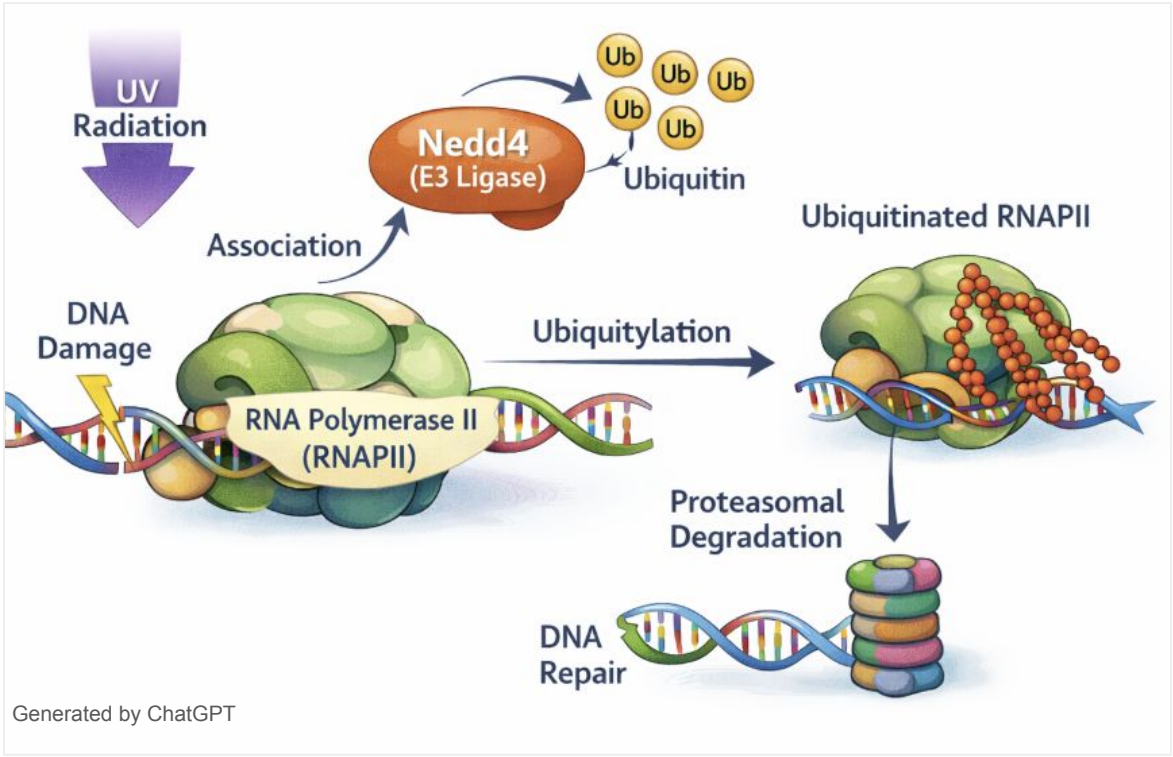


NEDD4 downregulates Na^{2+} channels



Generated by ChatGPT

NEDD4 inhibits gene transcription in the presence of DNA damage



Standard annotations are unconnected and lack context

Molecular Function

- ubiquitin-protein ligase activity
- sodium channel inhibitor activity
- RNA polymerase binding
- fibroblast growth factor binding

Cellular Component

- cytosol
- plasma membrane
- nucleus

Biological Process

- ubiquitin-dependent protein catabolic process
- positive regulation of endocytosis
- negative regulation of sodium ion transport
- cellular response to UV
- negative regulation of transcription from RNA polymerase II promoter
- negative regulation of fibroblast growth factor receptor signaling pathway

Dissecting the biology

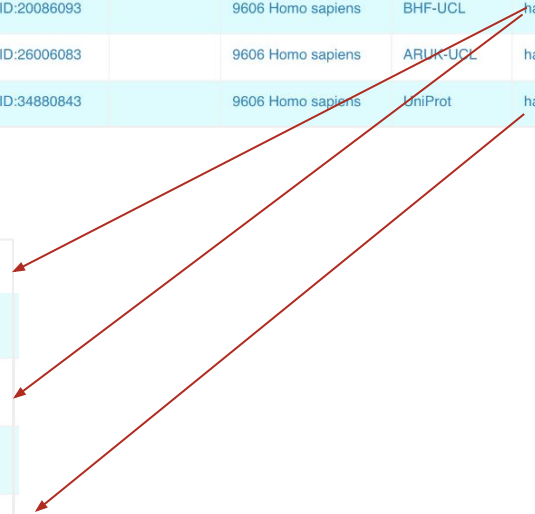
Molecular function (33 annotations + 75 'protein binding')

substrates

UniProtKB:P46934	NEDD4	enables	GO:0061630 F	ubiquitin protein ligase activity	ECO:0000314	IDA	PMID:17116753	9606 Homo sapiens	ARUK-UCL	has_input (UniProtKB:Q9NZ52)
UniProtKB:P46934	NEDD4	enables	GO:0061630 F	ubiquitin protein ligase activity	ECO:0000314	IDA	PMID:17996703	9606 Homo sapiens	BHF-UCL	part_of (GO:0000122) more...
UniProtKB:P46934	NEDD4	enables	GO:0061630 F	ubiquitin protein ligase activity	ECO:0000314	IDA	PMID:20086093	9606 Homo sapiens	BHF-UCL	has_input (UniProtKB:Q07912)
UniProtKB:P46934	NEDD4	enables	GO:0061630 F	ubiquitin protein ligase activity	ECO:0000314	IDA	PMID:26006083	9606 Homo sapiens	ARUK-UCL	has_input (UniProtKB:P08183) more...
UniProtKB:P46934	NEDD4	enables	GO:0061630 F	ubiquitin protein ligase activity	ECO:0000314	IDA	PMID:34880843	9606 Homo sapiens	UniProt	has_input (UniProtKB:Q7Z434) more...

Biological process (52 annotations)

UniProtKB:P46934	NEDD4	involved_in	GO:0000122 P	negative regulation of transcription by RNA polymerase II
UniProtKB:P46934	NEDD4	involved_in	GO:0006511 P	ubiquitin-dependent protein catabolic process
UniProtKB:P46934	NEDD4	involved_in	GO:0006974 P	DNA damage response
UniProtKB:P46934	NEDD4	involved_in	GO:0007041 P	lysosomal transport
UniProtKB:P46934	NEDD4	involved_in	GO:0010766 P	negative regulation of sodium ion transport
UniProtKB:P46934	NEDD4	involved_in	GO:0016567 P	protein ubiquitination



Early 2000's solution: Precomposition

Precomposed terms aimed to provide more expressive terms

cellular response to UV + negative regulation of transcription from RNA polymerase II promoter

=

negative regulation of transcription from RNA polymerase II promoter in response to UV-induced DNA damage

Examples obsolete pre-composed terms

- regulation of planar cell polarity pathway involved in axis elongation
- regulation of transcription involved in G1/S transition of mitotic cell cycle
- cell morphogenesis involved in differentiation
- regulation of the force of heart contraction involved in baroreceptor response to decreased systemic arterial blood pressure
- regulation of cytosolic calcium ion concentration involved in phospholipase C-activating G protein-coupled signaling pathway

Precomposition

cellular response to UV + negative regulation of transcription
from RNA polymerase II promot
response to UV-induced DNA damage

- Combinatorially unmanageable
- Complexity in the ontology leads to errors and inconsistencies
- Most have already been obsoleted

Annotation extensions

- AE provide additional context for annotations by linking MF, BP, CC annotations, as well as inputs/substrates, cell types and anatomical structures

Gene/product	Gene/product name	Annotation qualifier	GO class (direct)	Annotation extension
NEDD4	E3 ubiquitin-protein ligase NEDD4		ubiquitin protein ligase activity	part of negative regulation of transcription by RNA polymerase II part of DNA damage response

Huntley *et al.* 2014 BMC Bioinformatics

Annotation extensions

- AE provide additional context for annotation
- Limited scope because extensions are not easily exploitable by GO tools
- Not amenable to pathway-like representation

Annotation extensions ≠ pathways

Gene/product	Gene/product name	Annotation qualifier	GO class (direct)	Annotation extension
NEDD4	E3 ubiquitin-protein ligase NEDD4		ubiquitin protein ligase activity	part of negative regulation of transcription by RNA polymerase II part of DNA damage response

allowed extension

negatively regulates RNA polymerase activity

disallowed extension
(logic too complex for format)

negatively regulates RNA polymerase activity of **POL2A**

GO causal activity models (GO-CAMs)

Provides a system to extend GO annotations

✅ **provides context**

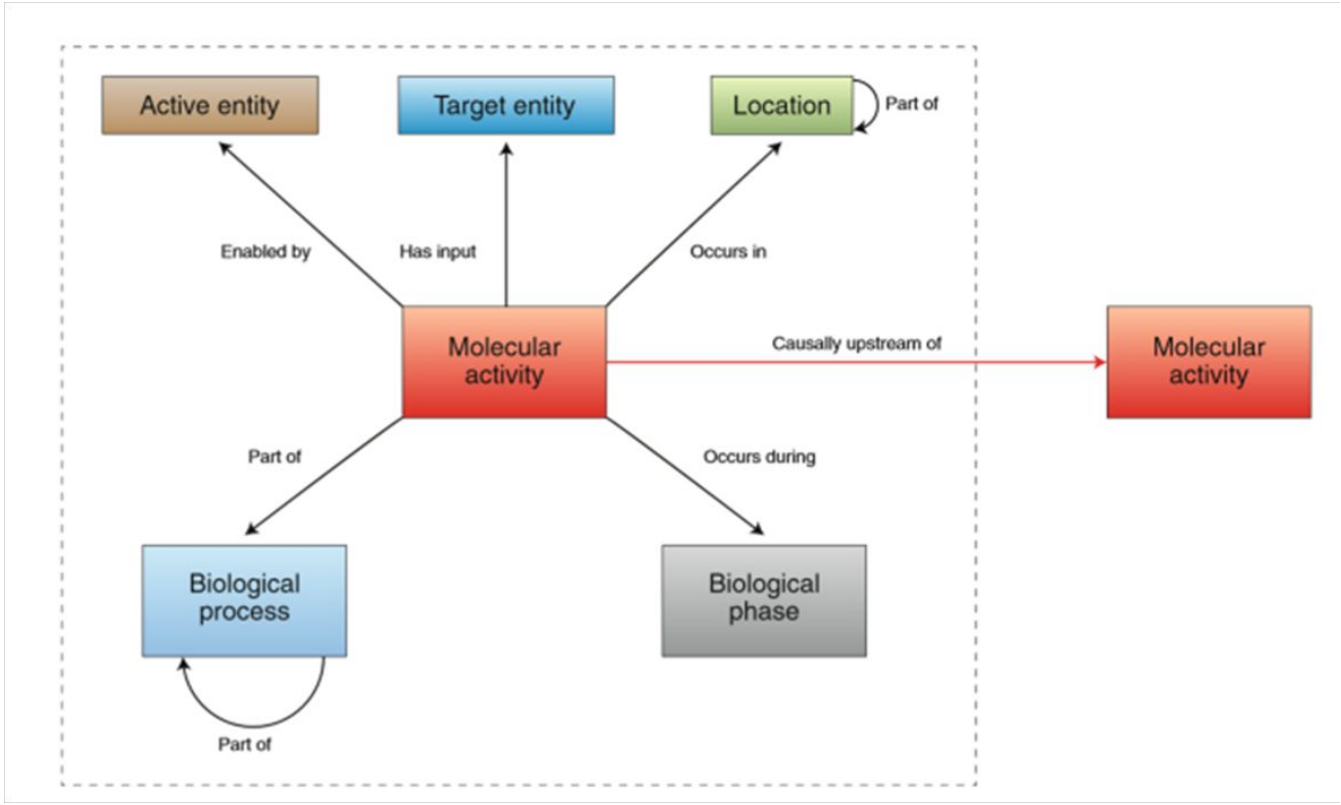
- CC: in which cellular component, cell, tissue the function/process take place
- MF: capture substrates
- BP: which BP is a MF part of
- BPs can also be part of broader BPs

✅ **defines causal connections between activities (MFs)**

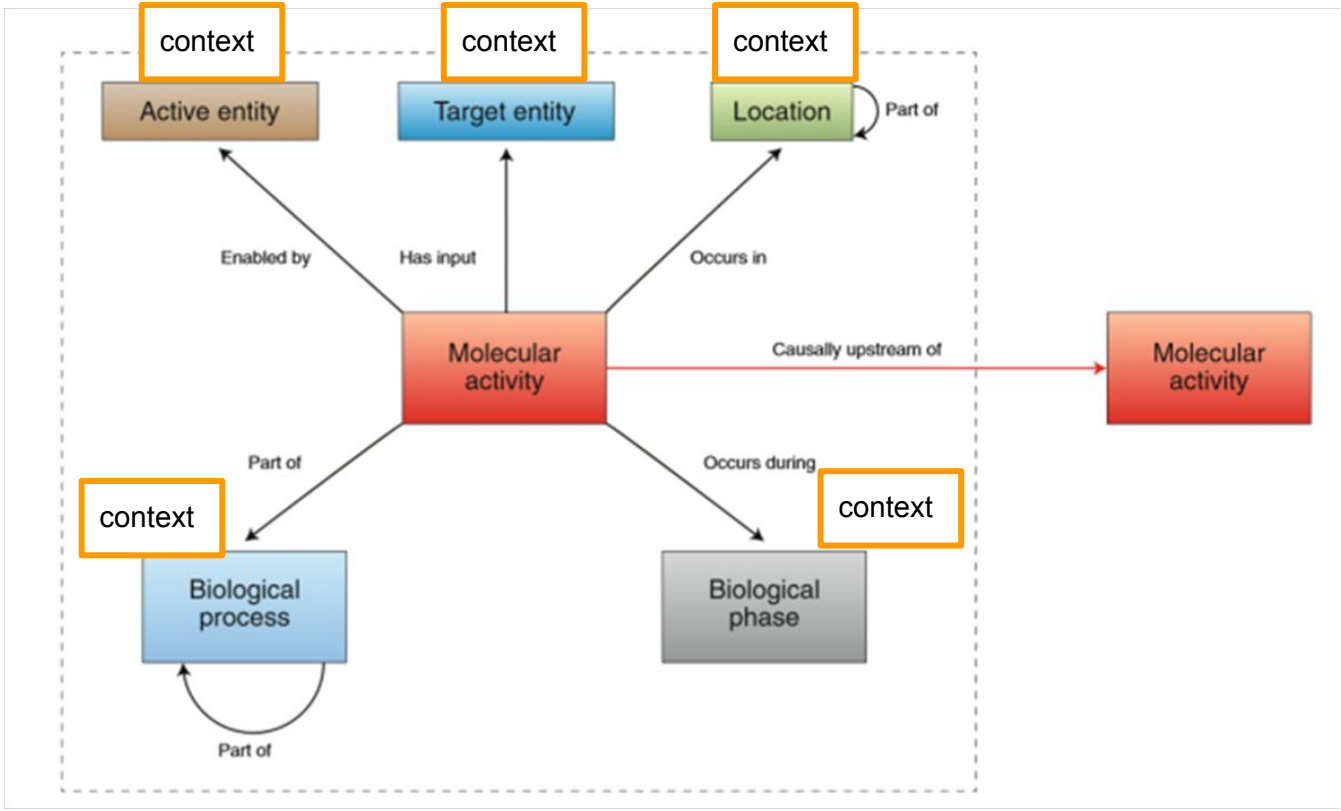
- allows network representation
- enables pathway visualization & analysis

Published in Thomas & al, Nat Genet. 2019
See also GO documentation

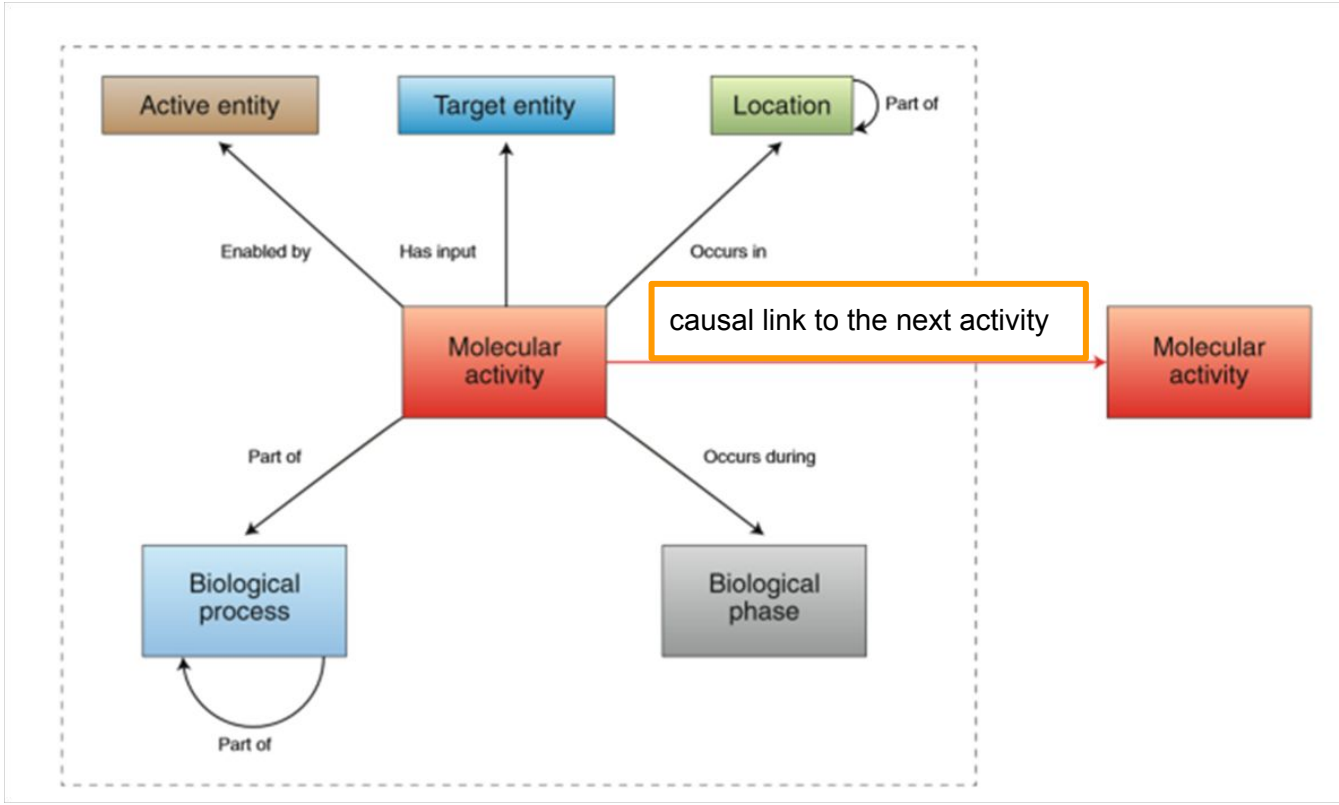
GO-CAM model specifications



GO-CAM model specifications



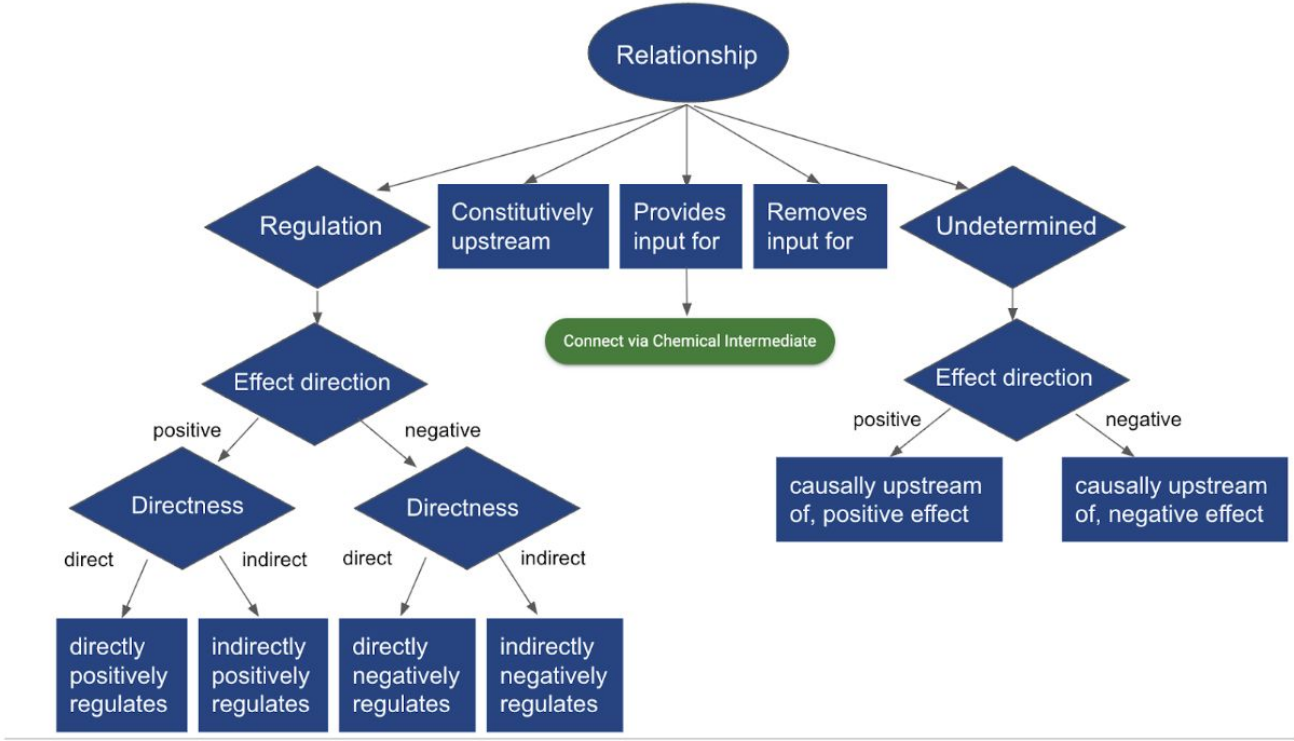
GO-CAM model specifications



Causal relations

- Link a MF to a MF
- Can be positive or negative

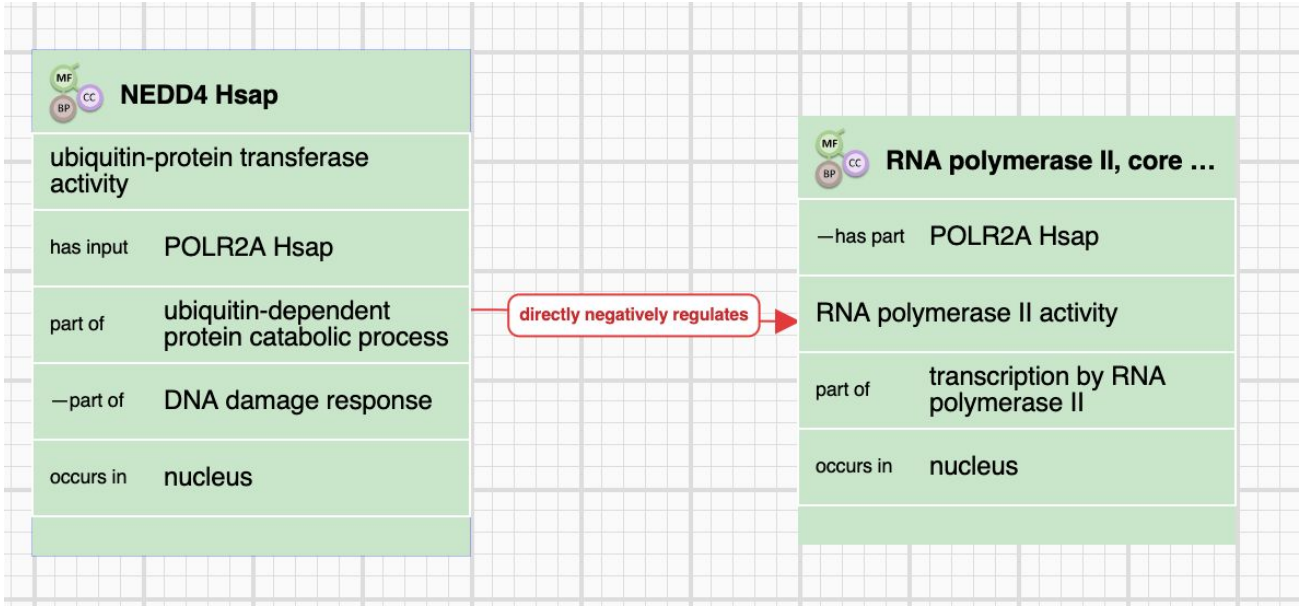
GO-CAM causal relations



See also the [Noctua User Guide]

directly negatively regulates	Links two activities when the upstream activity has a negative causal effect (decreasing or inhibiting) on an immediately downstream activity. Immediately means there is no intervening activity. The mechanism by which the upstream activity controls the downstream activity must be known.
directly positively regulates	Links two activities when the upstream activity has a positive causal effect (increasing or activating) on an immediately downstream activity.
indirectly negatively regulates	Links two activities when the upstream activity has a negative regulatory effect (decreasing or inhibiting) on the downstream activity via a larger process (e.g. proteasome-mediated protein degradation) that is reused in many contexts and the curator does not want to reproduce that process in the GO-CAM. The mechanism by which the upstream activity controls the downstream activity must be known.
indirectly positively regulates	Links two activities when the upstream activity has a positive regulatory effect (increasing or activating) on the downstream activity via a larger process (e.g. transcription) that is reused in many contexts and the curator does not want to reproduce that process in the GO-CAM. The mechanism by which the upstream activity controls the downstream activity must be known.
provides input for	Links two successive activities when the product (output) of the upstream activity is the substrate (input) for the downstream activity, and the product is a macromolecule (i.e. DNA, RNA, protein).
removes input for	Links two activities when the upstream activity has a negative causal effect (decreasing or inhibiting) on the downstream activity and the two activities act on or modify the same molecular target at the same site(s).
constitutively upstream of	Links two activities when the upstream activity is required for the downstream activity, but does not regulate the downstream activity.
causally upstream of, negative effect	Links two activities when the upstream activity has a negative causal effect (decreasing or inhibiting) on the downstream activity but the mechanism is not known.
causally upstream of, positive effect	Links two activities when the upstream activity has a positive causal effect (increasing or activating) on the downstream activity but the mechanism is not known.

GO-CAM model view fo regulation of RNA polymerase by NEDD4



GO-CAM impact on the ontology

Ontology refactoring

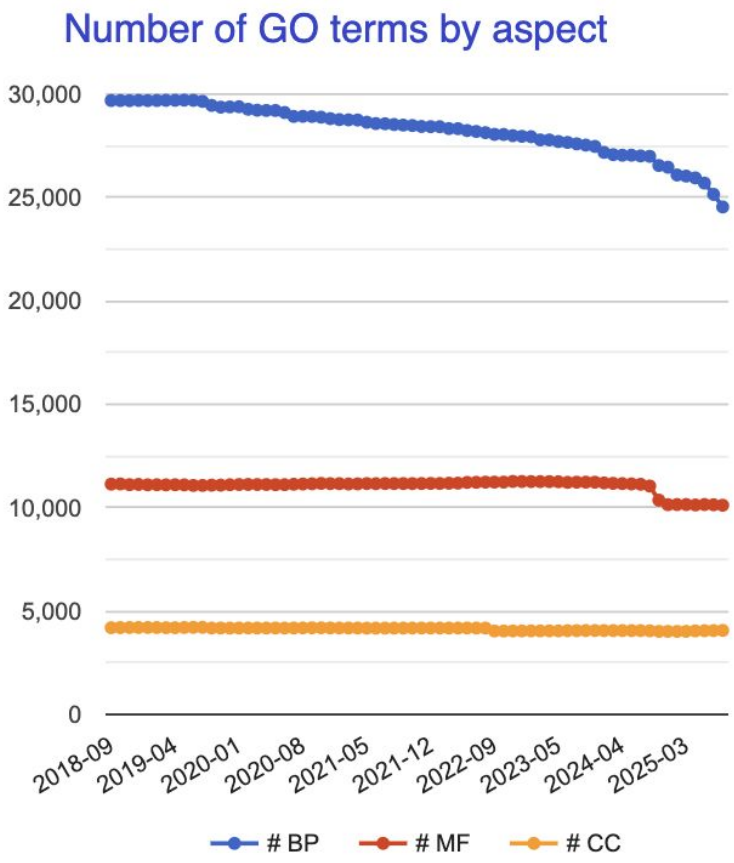
Started a major ontology refactor in 2019 to address issues caused by old ontology practices

- Proactive introduction of terms to anticipate potential curator needs (but never used)
- Excessive intermediate-level grouping terms
- Precomposed terms
- Single step reactions in the BP aspect of the ontology
- Terms that refer to more than one ontological aspect
- Experimental assays and non-physiological substrates

Examples obsoleted terms

Reason for obsoletion	Examples
Phenotype of perturbation experiments	*negative regulation of necrotic cell death* (GO:0060547)
BP terms that represent a MF	*histone H2A acetylation* (GO:0043968)
Pre-composed terms that are now represented by GO-CAMs	*histone H4 acetylation involved in response to DNA damage stimulus* (GO:2000776)
Sub-reaction and reaction mechanisms	*formation of peptidyl-cystine persulfide by sulphur transfer from free cysteine* (GO:0044526)
Substrates beyond the specificity of known enzymes or not physiologically relevant	*4,4′-diapophytoene desaturase (4,4′-diapolycopene-forming)* (GO:0140868)

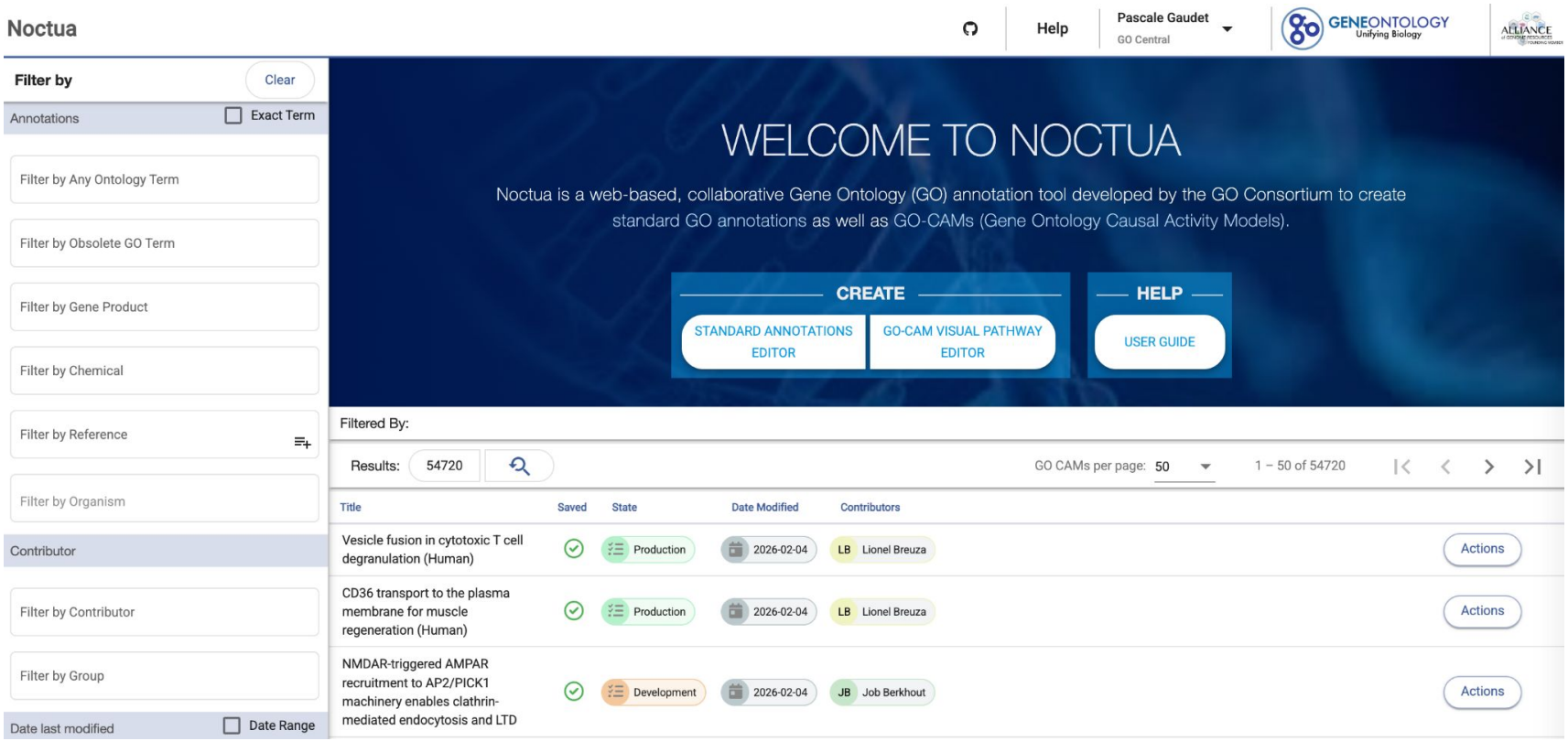
The Gene Ontology Consortium 2026, NAR Database issue


Number of GO terms by aspect
30,000
25,000
20,000
15,000
10,000
5,000
0
2018-09
2019-04
2020-01
2020-08
2021-05
2021-12
2022-09
2023-05
2024-04
2025-03
BP
MF
CC

WELCOME TO NOCTUA

Noctua is a web-based, collaborative Gene Ontology (GO) annotation tool developed by the GO Consortium to create standard GO annotations as well as GO-CAMs (Gene Ontology Causal Activity Models).

Noctua allows to view, create and edit GO annotations

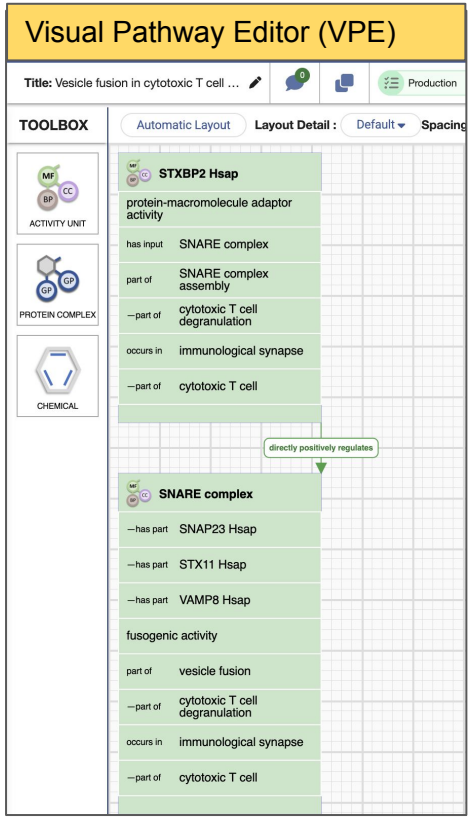


Two main workbenches

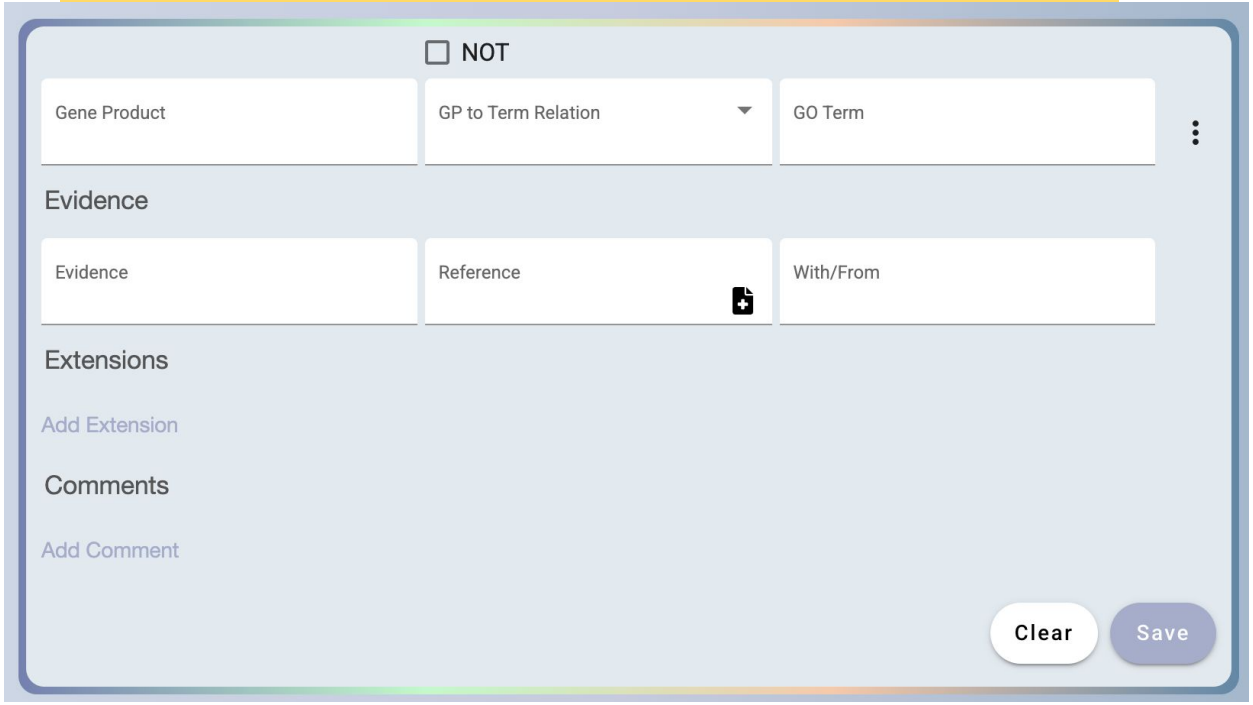
Standard Annotations Editor (SAE)

GENE PRODUCT	RELATIONSHIP	TERM	EVIDENCE	REFERENCE	WITH	EXTENSION	DATE MODIFIED
STXBP2 Hsap UniProtKB:Q15833	enables	protein-macromolecule adaptor activity GO:0030674	direct assay evidence used in manual assertion ECO:0000314	PMID:28265073		*has input* SNARE complex GO:0031201 *occurs in* immunological synapse GO:0001772 *part of* SNARE complex assembly GO:0035493	Feb 4, 2026
SNARE complex GO:0031201	enables	fusogenic activity GO:0140522	direct assay evidence used in manual assertion ECO:0000314	PMID:28265073		*occurs in* immunological synapse GO:0001772 *part of* vesicle fusion GO:0006906	Feb 4, 2026

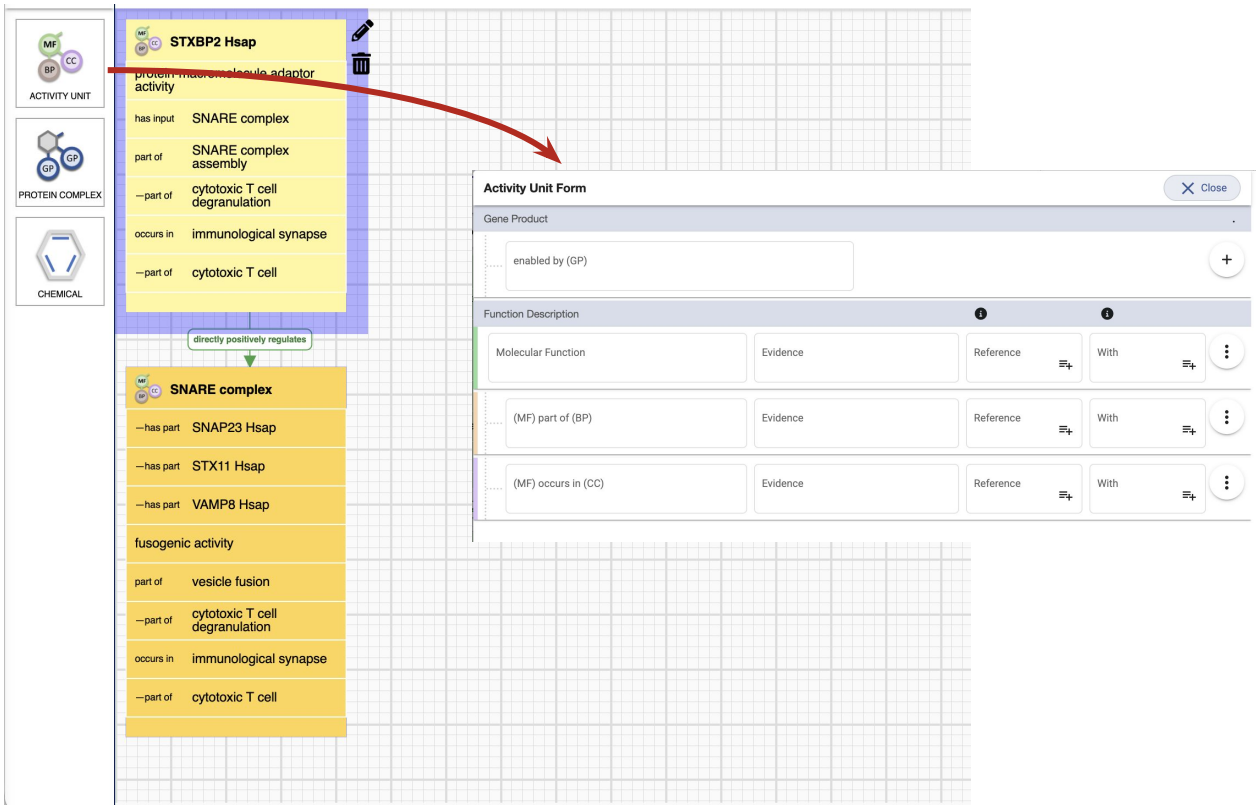
Visual Pathway Editor (VPE)



Standard Annotations Editor - creating annotations



Visual Pathway Editor - new annotation



GO-CAM status

- ~ 2,000 models ('production')
- > 10,000 distinct genes
- 5 major contributing groups: UniProt, PomBase, FlyBase, SGD, MGI

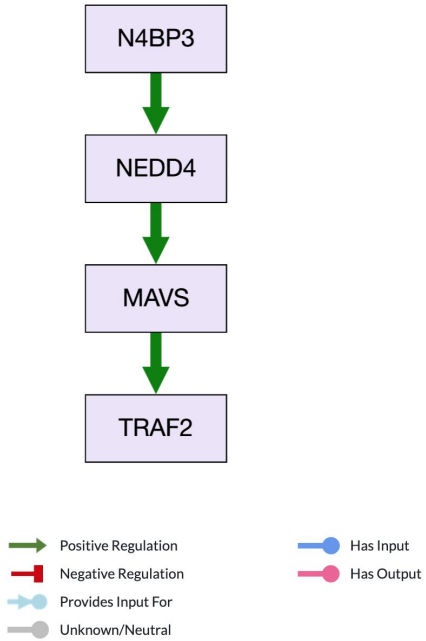
GO-CAM model availability

- GO-CAM model in OWL-type formats (ttl)
- conversion to standard annotations, with extensions if appropriate

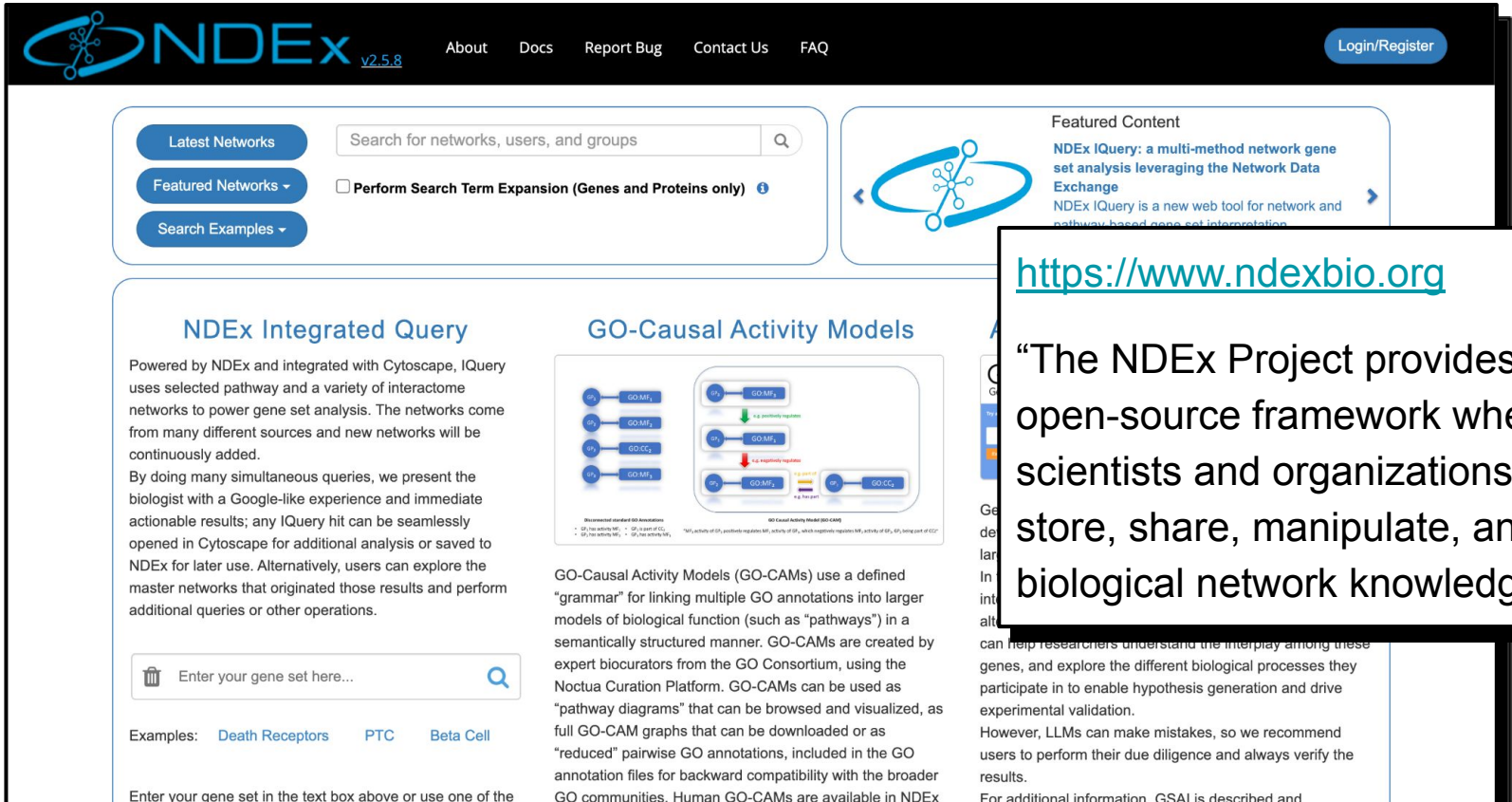
Gene/product	Gene/product name	Annotation qualifier	GO class (direct)	Annotation extension
NEDD4	E3 ubiquitin-protein ligase NEDD4		ubiquitin protein ligase activity	part of negative regulation of transcription by RNA polymerase II part of DNA damage response

GO-CAM Pathway Viewer widget

- Widget can be embedded in any website
- Available in AmiGO, Alliance of Genomes, UniProt, PomBase, FlyBase



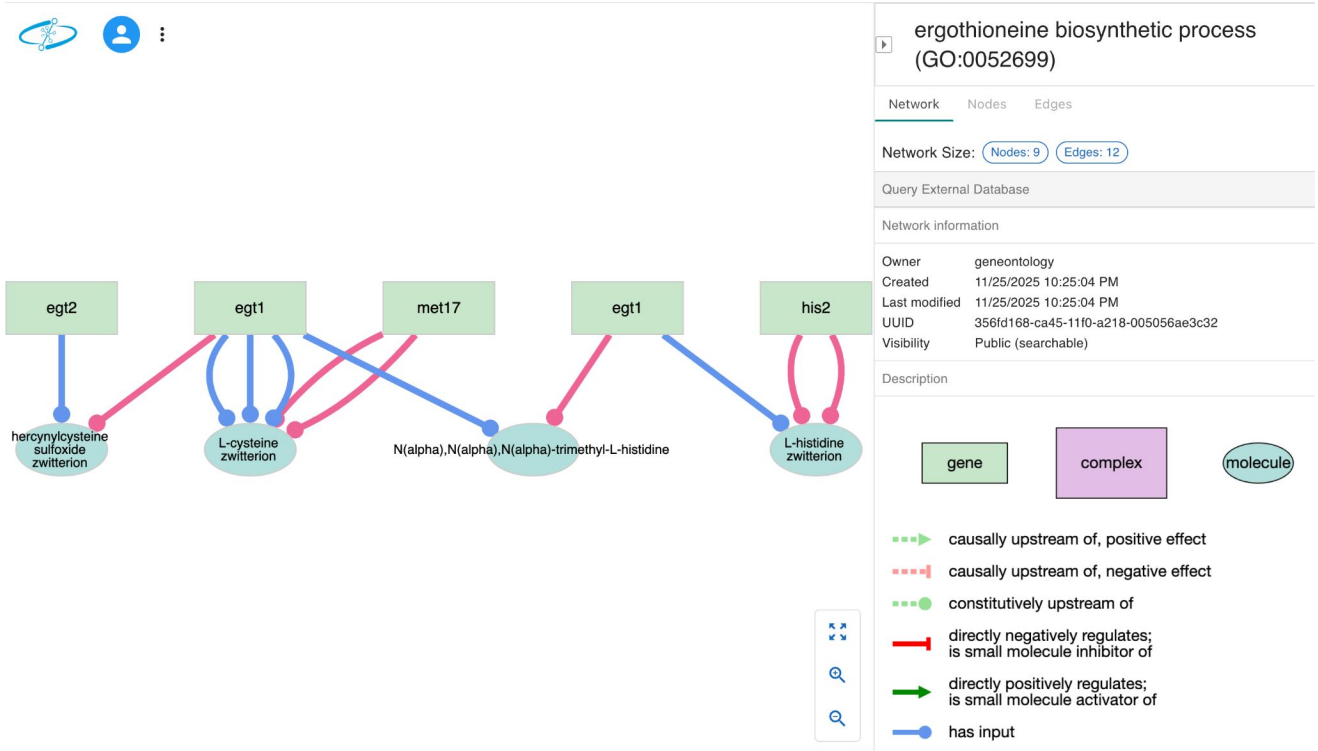
GO-CAMs are also available at NDEx



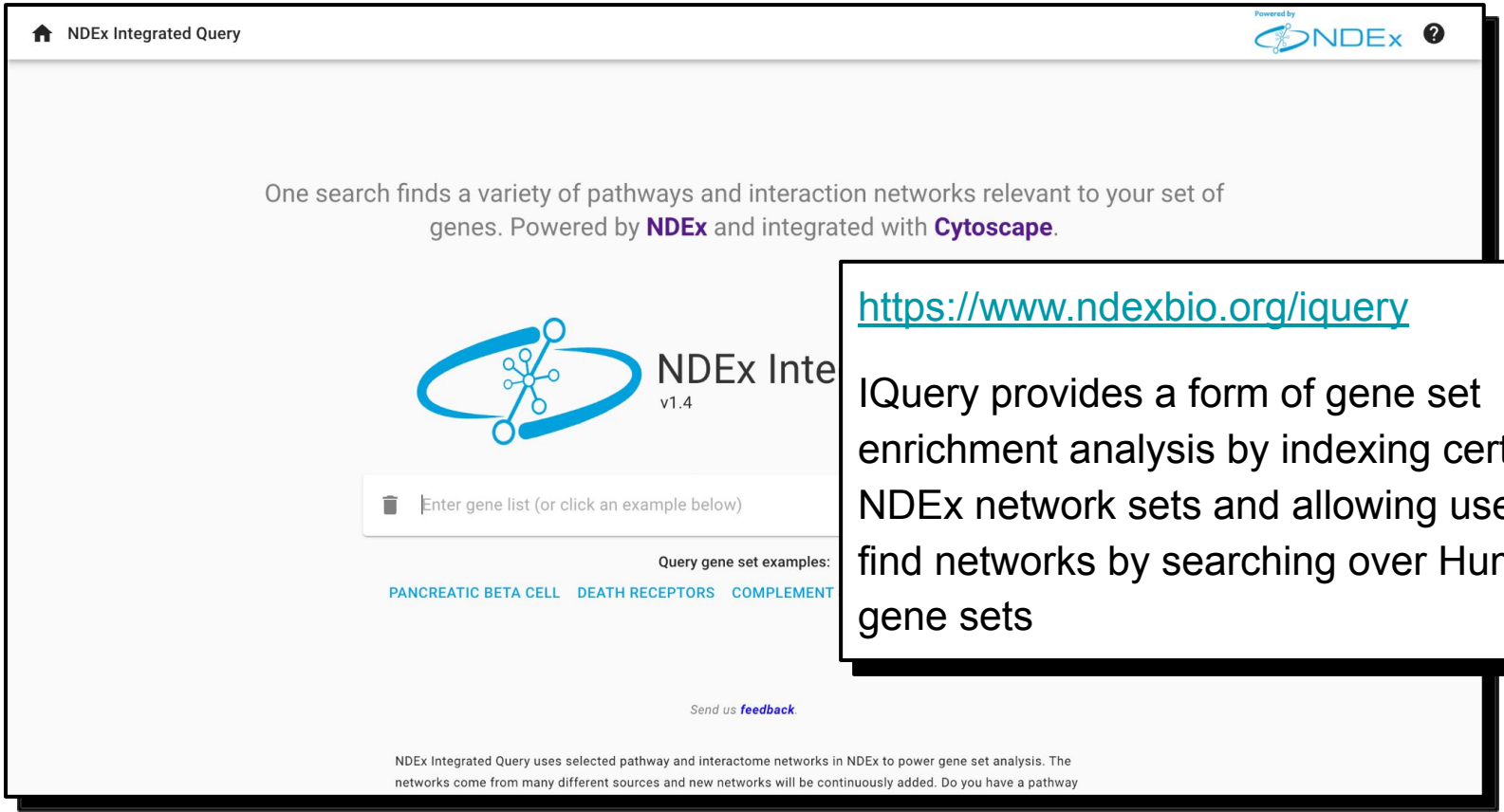
https://www.ndexbio.org

"The NDEx Project provides an open-source framework where scientists and organizations can store, share, manipulate, and publish biological network knowledge."

NDEx GO-CAM pathway viewer



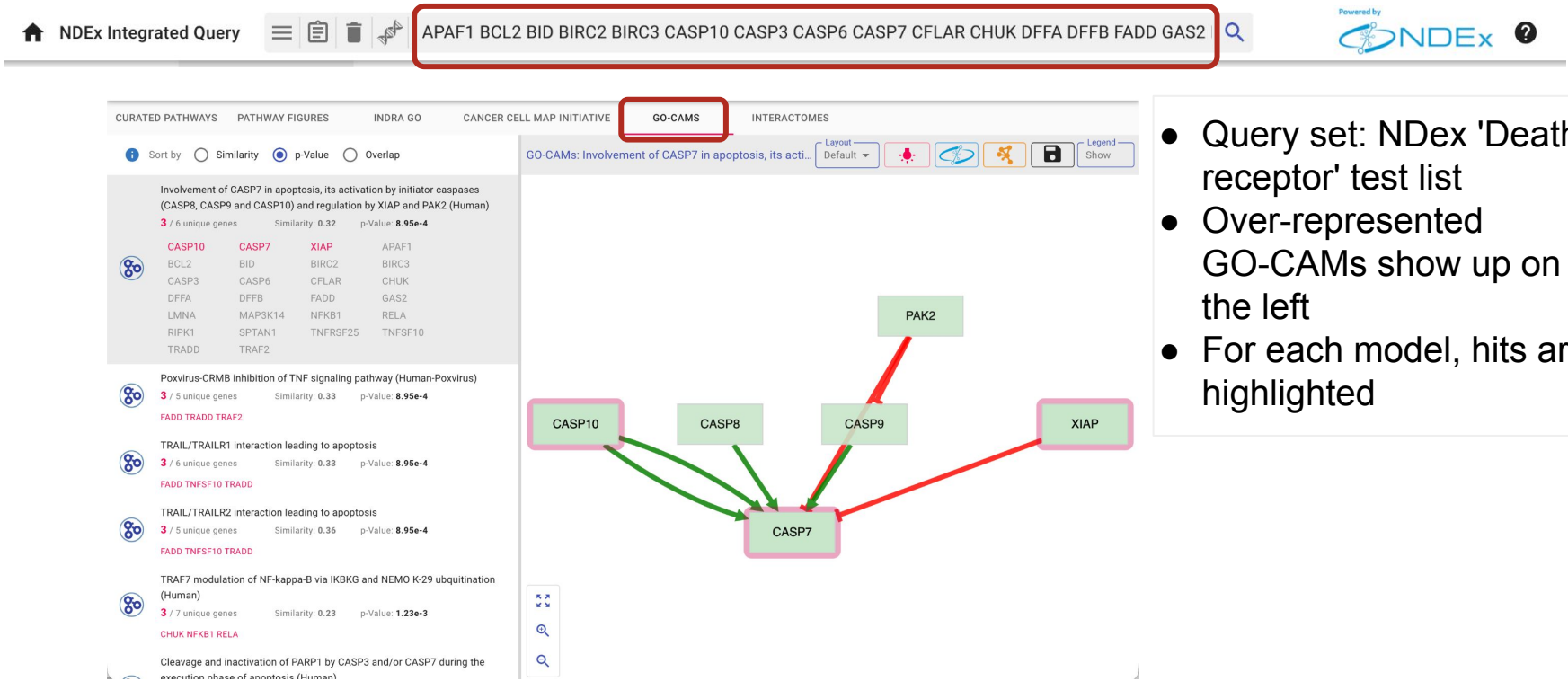
NDEx IQuery



https://www.ndexbio.org/iquery

IQuery provides a form of gene set enrichment analysis by indexing certain NDEx network sets and allowing users to find networks by searching over Human gene sets

NDEx IQuery tests pathway representation for lists of genes



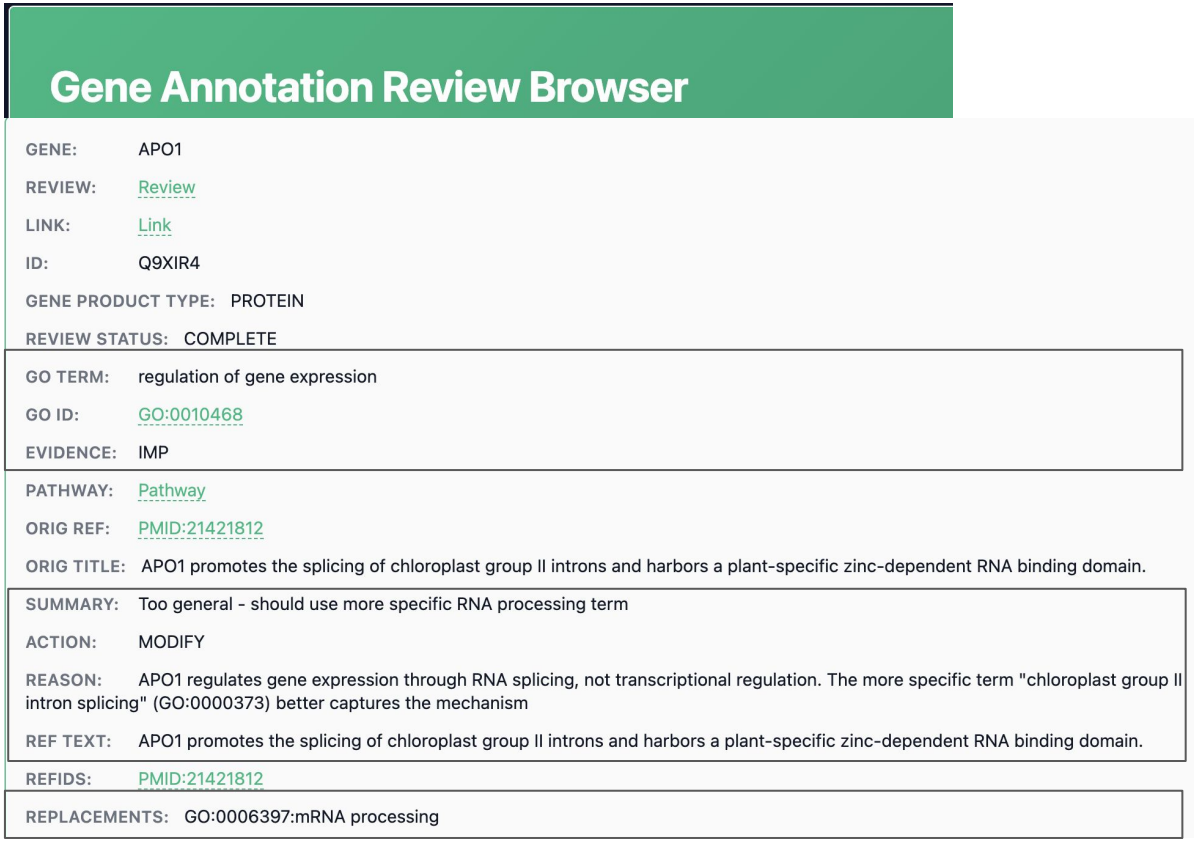
- Query set: NDex 'Death receptor' test list
- Over-represented GO-CAMs show up on the left
- For each model, hits are highlighted

Exploring AI to enhance GO

GO tasks

- Ontology development
  - create new terms, obsolete terms
  - fix hierarchy to align with external resources (EC, RHEA, etc)
  - apply design patterns
- Annotation
  - add new annotation
  - remove incorrect annotations
  - annotation reviews
  - new GO-CAMs
  - review GO-CAMs
- QC

AI-generated annotation review

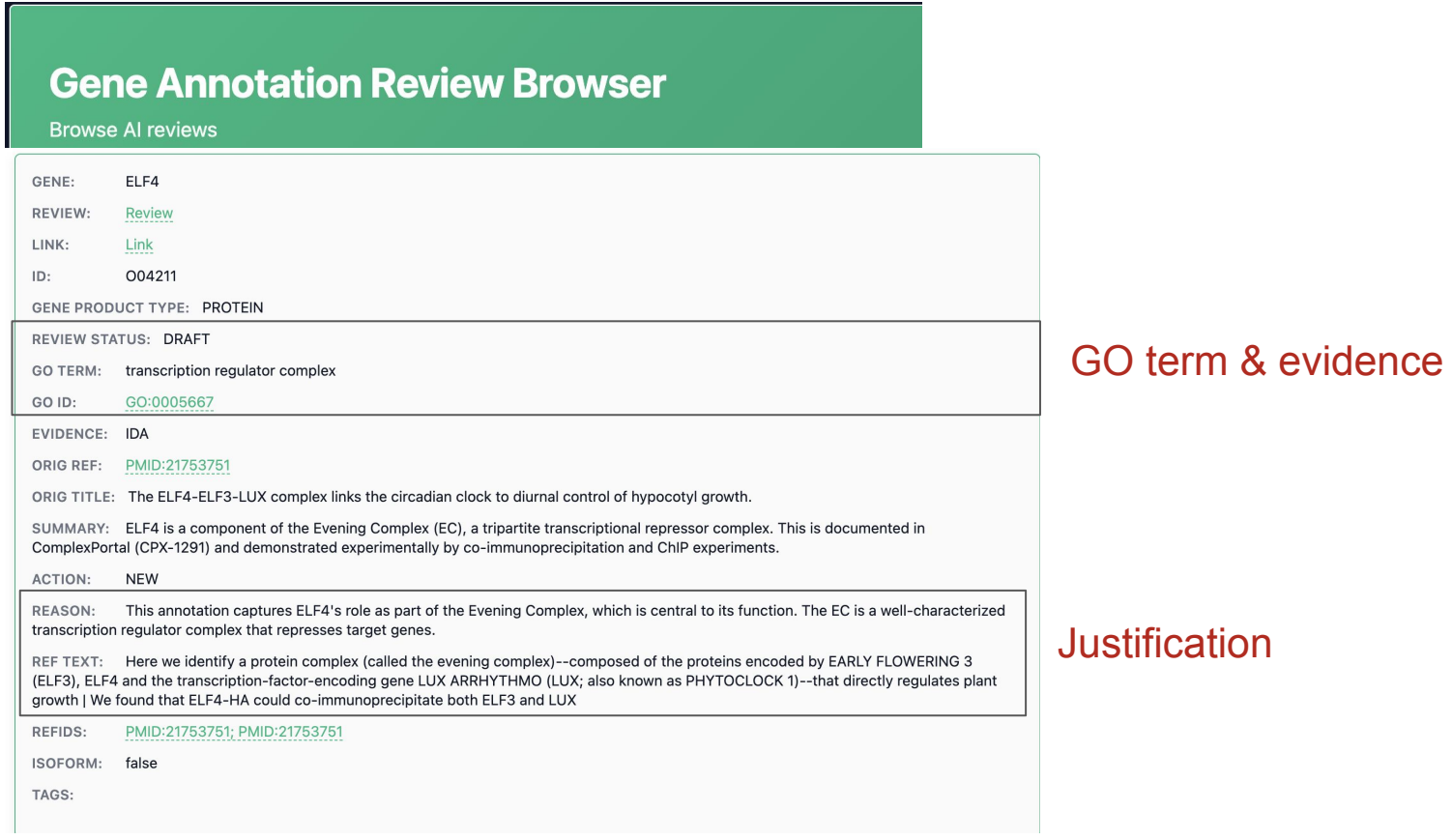


GO term & evidence

Justification

Suggested GO term

AI-generated new annotation suggestions

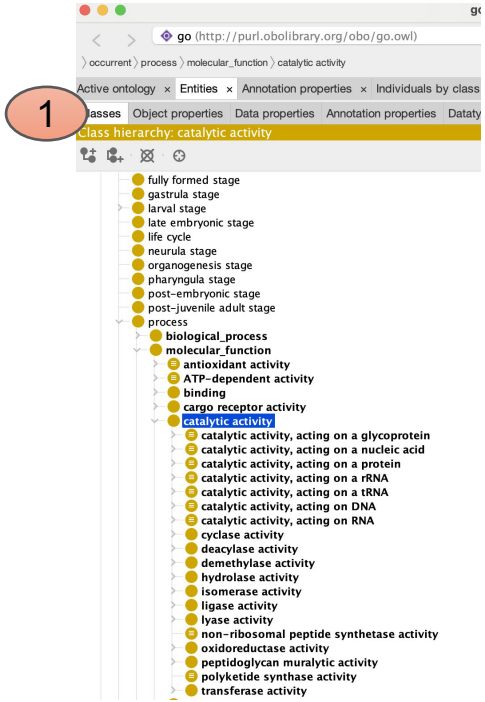


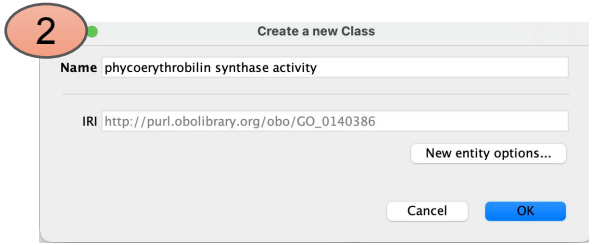
GO term & evidence

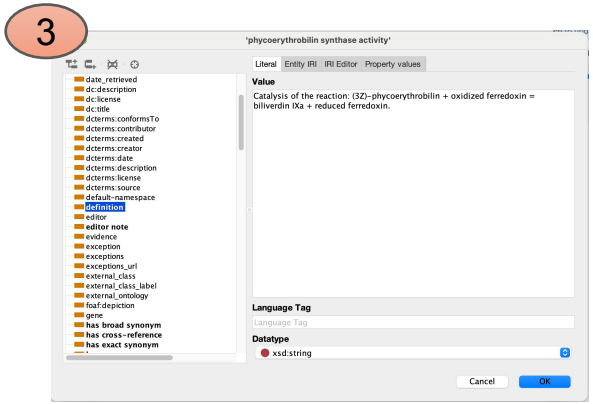
Justification

Conventional workflow for ontology editing

- GO uses Protégé as ontology editing tool, hooked to GitHub actions for QC







4 add namespace

5 add cross-reference #1

6 add cross-reference #2

7 add synonym

etc

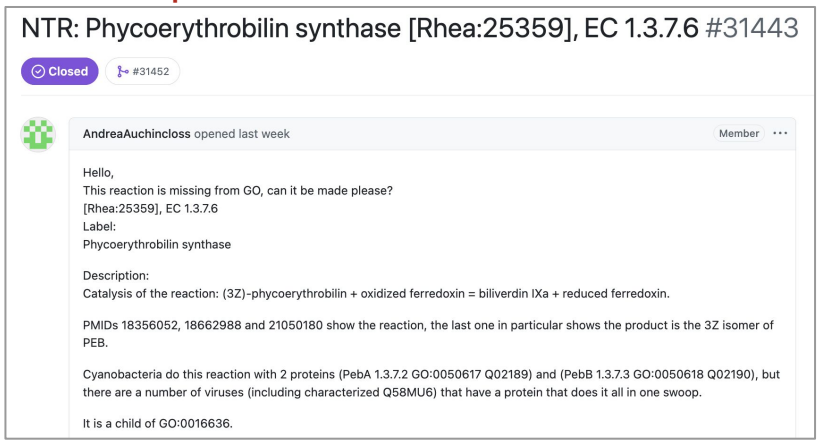
8 save

9 commit, push

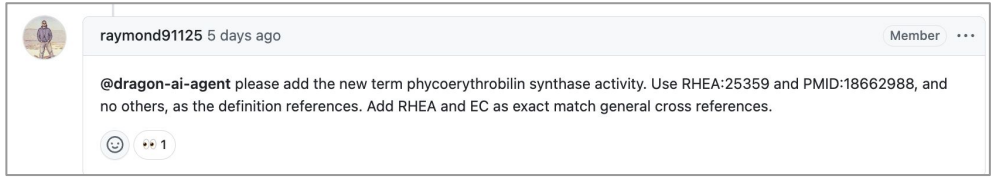
10 create pull request, merge

Agentic workflow

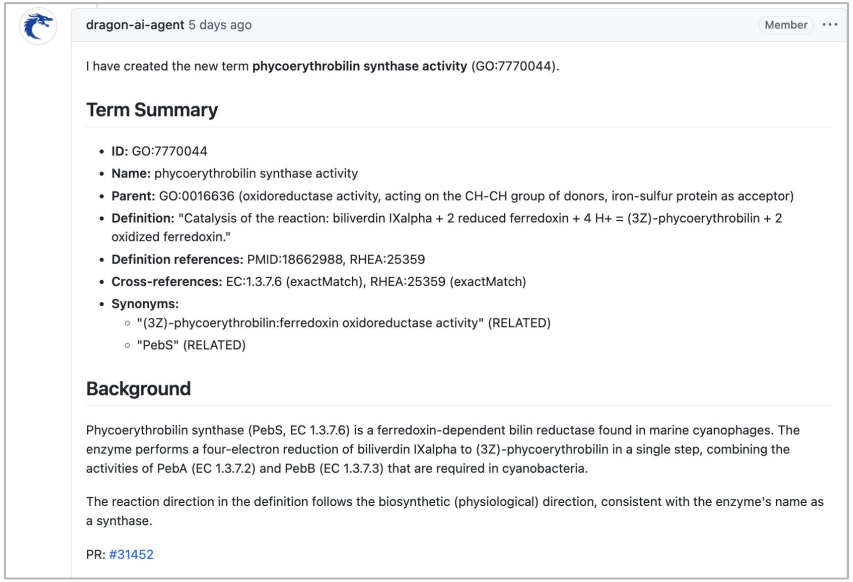
User request in GitHub



Editor asks agent to do the task



Agentic workflow - new term created !



dragon-ai-agent 5 days ago — Member

I have created the new term **phycoerythrobilin synthase activity** (GO:7770044).

Term Summary

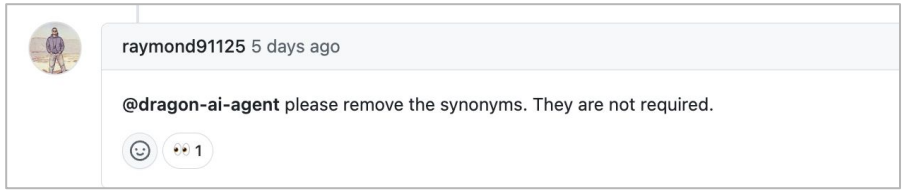
- **ID:** GO:7770044
- **Name:** phycoerythrobilin synthase activity
- **Parent:** GO:0016636 (oxidoreductase activity, acting on the CH-CH group of donors, iron-sulfur protein as acceptor)
- **Definition:** "Catalysis of the reaction: biliverdin IXalpha + 2 reduced ferredoxin + 4 H+ = (3Z)-phycoerythrobilin + 2 oxidized ferredoxin."
- **Definition references:** PMID:18662988, RHEA:25359
- **Cross-references:** EC:1.3.7.6 (exactMatch), RHEA:25359 (exactMatch)
- **Synonyms:**
  - "(3Z)-phycoerythrobilin:ferredoxin oxidoreductase activity" (RELATED)
  - "PebS" (RELATED)

Background

Phycoerythrobilin synthase (PebS, EC 1.3.7.6) is a ferredoxin-dependent bilin reductase found in marine cyanophages. The enzyme performs a four-electron reduction of biliverdin IXalpha to (3Z)-phycoerythrobilin in a single step, combining the activities of PebA (EC 1.3.7.2) and PebB (EC 1.3.7.3) that are required in cyanobacteria.

The reaction direction in the definition follows the biosynthetic (physiological) direction, consistent with the enzyme's name as a synthase.

PR: #31452



raymond91125 5 days ago

@dragon-ai-agent please remove the synonyms. They are not required.

👀 1

Editor asks agent to make corrections if needed

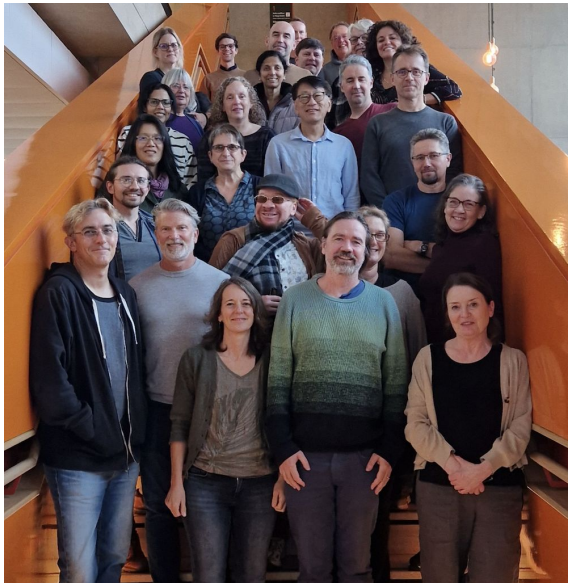
Summary

- GO is a computational model of gene function that aims to be comprehensive, species-agnostic and current
- The ontology is continually updated, with recent major updates to better capture biology rather than experimental results
- Annotations exist in both standard and GO-CAM flavors
- GO-CAMs are now disseminated across various resources
- GO strives to be efficient and accurate

Ressources

- GO Handbook
- GO-CAM paper (2019)
- Gene Ontology website
  - General ontology documentation
  - General annotation documentation
- Annotation Relations documentation
- EBI Gene Ontology tutorial on YouTube (general)

Contributors



PIs
Paul Thomas, Chris Mungall, Paul Sternberg

GO Central
Suzi A. Aleksander, James Balhoff, Seth Carbon, Dustin Ebert, Marc Feuermann, Raymond Lee, Huaiyu Mi , Sierra Moxon, Anushya Muruganugan Tremayne Mushayahama, Kimberly Van Auken

Contributing groups
CGD, CACAO, EcoWiki, ComplexPortal, dictyBase, DisProt, EcoCyc, EcoliWiki, Ensembl, FlyBase, HGCN HUGO, IntAct, InterPro, JaponicusDB, MGI, PANTHER, PHI-base, PomBase, RGD, Reactome, RHEA, RNACentral, SGD, SignatureScience, TAIR, HPA, Matrisome, SynGO, UniProt, WormBase, ZFIN