

Chemical Data Extractor

User Tutorial

From your first search to advanced, repeatable workflows

Version 3.2 · Complete guide, beginner to advanced

19 database connectors · multi-hop graph traversal · enrichment pipeline

Covers: concepts, accounts, running a search, reading a run, export files, search history, the AI analyst, plans and payment, light and dark mode, troubleshooting, and batch and command-line use.

Generated from the application's documentation source (`docs/tutorial/user-tutorial.md`), so it always matches the shipped interface.

Contents

What you will learn	4
Part 1 — Concepts	4
1.1 What a knowledge graph is here	4
1.2 What "hops" means	4
1.3 What the engine does, in order	5
Part 2 — Your account	5
2.1 Creating an account	5
2.2 Signing in later	6
2.3 Where your plan and usage are shown	6
2.4 Plans and limits	6
Part 3 — Running a search	6
3.1 The search form	7
3.2 Step-by-step: your first search	7
3.3 Good queries, and queries that do not work	7
Part 4 — Reading a run	8
4.1 Status	8
4.2 Progress	8
4.3 The engine log	9
4.4 Runtime expectations	9
Part 5 — Export files	9
5.1 What you get	9
5.2 Downloading	10
5.3 Which file to use	10
Part 6 — Search history	10
Part 7 — The AI analyst	11
Part 8 — Plans and payment	11
8.1 Buying a plan	11
8.2 The Academic plan	11
Part 9 — The interface	12
9.1 Light and dark mode	12
9.2 Keyboard and screen readers	12
9.3 Mobile and tablet	12
Part 10 — Troubleshooting	12
Part 11 — Reference	13

11.1 Data sources	13
11.2 Where results come from	14
11.3 Glossary	14
11.4 Where to get help	14

Chemical Data Extractor turns a single protein name or compound name into a multi-hop knowledge graph. It queries a set of public scientific databases, resolves the results into entities and relationships, and exports the whole graph as an Excel workbook, CSV, GraphML, Cypher, RDF, Parquet and more.

This guide takes you from your first search to advanced, repeatable workflows. No programming is required for anything in Part A to Part D.

What you will learn

Part	Contents	Who it is for
1. Concepts	Nodes, edges, hops, entities, relations	Everyone — read this first
2. Your account	Creating an account, plans, limits	Everyone
3. Running a search	Query, query type, expansion depth	Everyone
4. Reading a run	Status, progress, the engine log, errors	Everyone
5. Export files	What each download contains, retention	Everyone
6. Search history	Finding and re-opening earlier work	Everyone
7. The AI analyst	Asking questions about your results	Beginners to intermediate
8. Plans and payment	Buying a plan, academic verification	Anyone who outgrows the free tier
9. Interface	Light and dark mode, keyboard use, mobile	Everyone
10. Troubleshooting	Every error message explained	Everyone
11. Reference	Data sources, glossary, where results come from	Reference

Part 1 — Concepts

1.1 What a knowledge graph is here

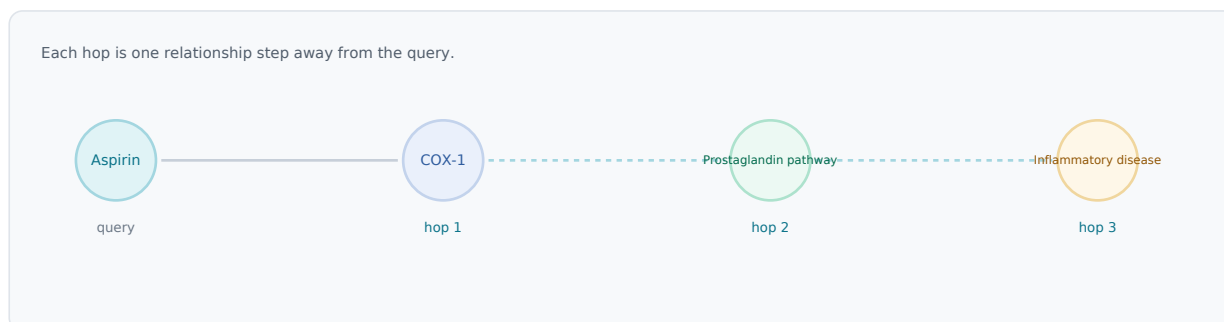
A search in this application produces two kinds of object.

- A **node** is a thing: a compound, a protein, a pathway, a disease, an organism, a publication or a patent. Every node carries identifiers from the databases it was found in.
- An **edge** is a relationship between two nodes, and it is typed. Typical types are INHIBITS, BINDS, HAS_STRUCTURE, PART_OF_PATHWAY and ASSOCIATED_WITH.

For example, a 2-hop search for *Aspirin* might produce the edge chain Aspirin –INHIBITS→ COX-1 –PART_OF_PATHWAY→ Arachidonic acid metabolism.

1.2 What "hops" means

A hop is one relationship step away from your starting point.



Multi-hop expansion diagram

- **1 hop** — only direct relationships. Fast, small, precise.

- **2 hops** — direct relationships plus the neighbours of those neighbours. This is the usual choice for compound screening.
- **3 hops** — adds pathway- and disease-level context. Larger exports.
- **4 hops** — maximum depth across every database. Broadest and slowest; expect minutes rather than seconds.

Each extra hop roughly multiplies the number of nodes, so a 4-hop search can produce tens of thousands of rows. Start at 1 or 2 hops and increase only when you need the extra context.

1.3 What the engine does, in order

1. **Routing and expansion** — detects the query type and walks outward through

every reachable database for the number of hops you chose.

1. **Entity resolution** — merges the same real-world thing found under several names and identifiers into a single node.

1. **Deduplication** — removes duplicate records left by overlapping databases.

2. **Enrichment** — fills in missing SMILES, molecular formula, molecular weight, titles and citations.

1. **Ranking** — scores entities by a combination of potency (pChEMBL), clinical phase, availability of an experimental structure, and how many independent databases corroborate them.

1. **Export** — writes the graph to 15 files across 7 formats.

You can watch all six steps in the engine log while a search runs.

Part 2 — Your account

2.1 Creating an account

Accounts are required to run searches. The free account needs nothing but an email address.

1. Open the application. Click **Sign in** in the top-right corner.
2. Choose **Create account** on the segmented control.
3. Enter your name (optional), email address, and a password of at least eight

characters.

1. Click **Create account**.
2. If your deployment requires email confirmation, open the confirmation email

and click the link, then sign in.

Your account ×

Create a free account to start searching.

Sign in Create account

EMAIL

you@university.edu

PASSWORD

At least 8 characters

Create account

[Forgot password?](#)

No card needed.
Three free searches are waiting.

The account dialog with the Create account tab selected

2.2 Signing in later

Use **Sign in**, enter your email and password, and you are done. If you have forgotten your password, choose **Forgot password?** and enter your email address. A reset link is sent to that address; opening it takes you to a page where you can set a new password.

For security, the same confirmation message is shown whether or not an account exists for the address you entered.

2.3 Where your plan and usage are shown

Your plan name and usage appear in the top-right of the application, for example Free · 2/3. Opening it shows:

- searches used against your plan allowance, and the AI analyst allowance;
- your academic profile, which you can edit and save at any time;
- **Upgrade plan** and **Sign out**.

2.4 Plans and limits

Plan	Price / month	Searches	Analyst messages
Free	₹0	3	5
Academic	₹249	30	100
Starter	₹499	50	250
Professional	₹1,999	300	1,500
Team	₹7,499	3,000	10,000

Two operational limits are worth knowing:

- **One search at a time.** If a search is already running, yours is queued and starts as soon as the first one finishes. Nothing is lost.
- **Exports are temporary.** Generated files are deleted from the server about 30 minutes after the search completes. Download anything you need to keep.

Part 3 — Running a search

3.1 The search form

The search form with the query field, query type and expansion depth

The form has three controls and one button.

Query. The name of the protein or compound you want to explore, for example tubulin, EGFR, Aspirin or Ibuprofen. A specific identifier such as P23219 works too and is usually more precise than a common name.

Query type.

Setting	Use it when
Auto	You are not sure. The engine detects protein vs compound.
Protein	The query is definitely a protein, gene or target.
Ligand	The query is definitely a small molecule or drug.

Auto is the right default. Choose Protein or Ligand only when you know the answer and Auto has got it wrong.

Expansion depth. The number of hops. The panel under the control explains what each depth buys you in one line.

Run search. Starts the search and locks the form until the run finishes.

3.2 Step-by-step: your first search

1. Click into the **Query** field and type tubulin.
2. Leave **Query type** on **Auto**.
3. Leave **Expansion depth** on **1**.
4. Click **Run search**.
5. Watch the right-hand panel. The status changes from *Queued* to *Running*, and the engine log begins to fill with live output.

1. When the status becomes **Completed**, the **Export files** panel appears above the log. Download what you need.

A 1-hop protein search normally finishes in well under a minute once the engine is awake.

3.3 Good queries, and queries that do not work

Query	Result
Aspirin	Excellent — very high coverage across chemistry databases.
EGFR	Good — but epidermal growth factor receptor is more precise.
tubulin	Good, broad. Expect many isoforms.
P23219	Best form for a protein — an unambiguous UniProt accession.
a drug for cancer	Does not work. The field takes a single entity name.
Aspirin, Ibuprofen	Batch input belongs in the CLI (Part 11), not the search box.

If a search returns almost nothing, the usual cause is a name that no database recognises. Try the alternative spelling, the full name, or a database identifier.

Part 4 — Reading a run

The screenshot shows a results panel for a completed run. At the top, a green pill indicates 'Completed' status, followed by the query 'tubulin · 1-hop · 33.8s' and a unique ID 'a3f19c7e'. Below this, a message states 'Pipeline finished - 14 export files written'. The 'Export files' section lists three files: 'tubulin_Knowledge_Graph.xlsx' (Excel workbook), 'nodes.csv' (CSV table), and 'graph.graphml' (GraphML graph). The 'ENGINE LOG' section shows a live stream of messages: '[1/6] Intelligent Routing & Expansion', '[OK] UniProt - 42 entities', '[OK] ChEMBL - 128 bioactivities', '[4/6] Running Graph Analytics Engine', '[warn] AlphaFold DB timed out - skipped', '[6/6] Generating multi-format exports', and '[OK] Pipeline finished successfully'.

The results panel showing status, progress, the engine log and export files

4.1 Status

The status pill in the top-left of the results panel is always one of:

Status	Meaning	What to do
Queued	The job is accepted but not yet running	Wait. See the note below.
Running	The engine is working	Wait; watch the log.
Completed	The graph and exports are ready	Open the Export files panel.
Failed	The run stopped	Read the message, then Part 10.

Two "Queued" situations look identical but are not. When the engine has been idle it wakes on your first request, which can take up to a minute — the panel says so. When another search is already running, your job waits for the single execution slot and starts by itself; the panel also says so explicitly.

4.2 Progress

A thin moving bar under the header shows that work is happening. The line under it names the current stage, for example Running expansion pass 2/4.

There is deliberately no percentage. The engine does not know the total amount of work in advance, and a fake percentage that freezes at 40% is worse than an honest moving bar.

4.3 The engine log

The log is the complete, unedited output of the search process. It is colour-coded by meaning:

Colour	Meaning
Grey	Ordinary progress output
Green	A step completed successfully ([OK])
Blue	The current pipeline phase, [3/6] and similar
Amber	A warning — often a database that timed out and was skipped
Red	An error
Box-drawing characters	The section banners the engine prints

Scrolling up pauses auto-follow. Scroll back to the bottom to resume.

4.4 Runtime expectations

Search	Typical runtime
1-hop compound	20–60 seconds
1-hop protein	30–90 seconds
2-hop	1–4 minutes
3-hop	3–10 minutes
4-hop	5–20 minutes

The first search after a period of inactivity is slower because the engine has to start. Exports are produced at the end of the run, so a long search appears idle during its final step — the log still shows progress.

Part 5 — Export files

Export files

14 files · 33.8s

xlsx

tubulin_Knowledge_Graph.xlsx

Excel workbook · all sheets

↓

csv

nodes.csv

One row per entity

↓

csv

edges.csv

One row per relationship

↓

cyp

neo4j_import.cypher

CREATE statements for Neo4j

↓

gra

graph.graphml

Whole graph for Gephi or Cytoscape

↓

ttl

graph.ttl

RDF triples for a linked-data store

↓

Exports are temporary — the server deletes generated files about 30 minutes after the search completes. Download what you keep.

The Export files panel listing the generated files

5.1 What you get

A completed search produces up to 15 files:

File	Type	Contains
------	------	----------

<query>_Knowledge_Graph.xlsx	Excel workbook	The main multi-sheet workbook: nodes, edges, bioactivity, summary
nodes.csv	CSV	One row per entity
edges.csv	CSV	One row per relationship
graph.graphml	GraphML	The complete graph for Gephi, Cytoscape or yEd
neo4j_import.cypher	Cypher	CREATE statements that load the graph into Neo4j
graph.ttl	Turtle	RDF triples for Blazegraph, Jena or any RDF store
rdf_graph.json	JSON	RDF-ready dictionaries, easy to post-process
nodes.parquet	Parquet	Columnar entity table for large analyses
edges.parquet	Parquet	Columnar relationship table
vector_index_metadata.json	JSON	Metadata for building an embedding index
triples.json	JSON	Flat subject-predicate-object triples
search_library.json	JSON	The resolved query terms and routing decisions
enriched_data.xlsx	Excel workbook	Enriched compounds, properties and publications
structure_images/	Folder	Rendered 2D structures, one image per compound

The exact list for your search is always shown in the interface — the table above is the usual set.

5.2 Downloading

Click any row in the **Export files** panel. The file opens in a new tab and your browser saves it. Downloading is per-file, so take the ones you need.

Downloads are removed from the server about 30 minutes after the search finishes. If you need to keep a graph, download it straight away. Re-running a search regenerates the files.

5.3 Which file to use

- **Spreadsheet analysis or a paper supplement** → the Excel workbook.
- **Statistics, pandas, R, or a database import** → nodes.csv and edges.csv.
- **Visualising the network** → graph.graphml, opened in Gephi or Cytoscape.
- **Your own graph database** → neo4j_import.cypher, which loads straight into Neo4j.
- **Linked data and RDF work** → graph.ttl.
- **Very large graphs** → the Parquet files.

Part 6 — Search history

The **History** tab lists your 20 most recent searches, newest first, with the status, the runtime and the number of export files. Clicking any row loads that search's status, log and progress into the workspace.

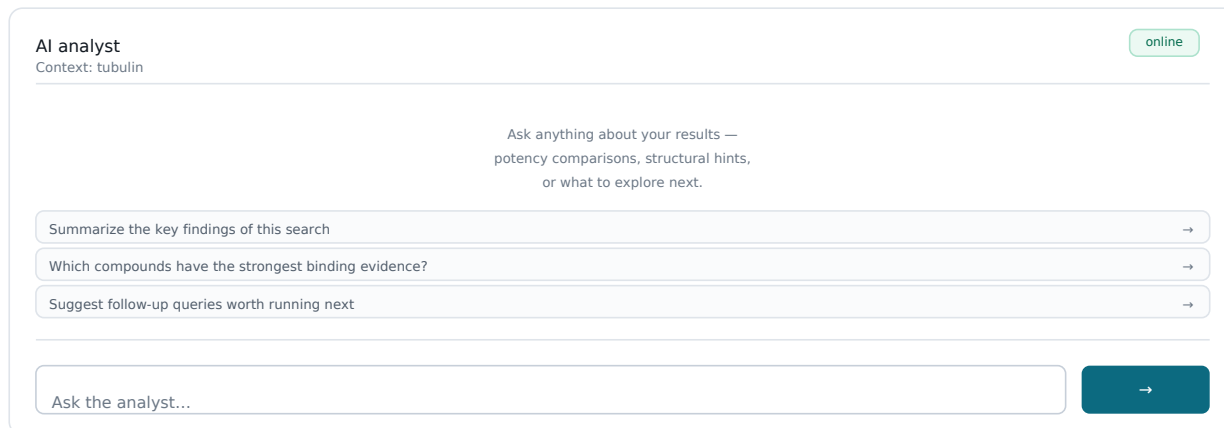
G Chemical Data Extractor		Search	History	Resources
Search history		20 most recent, newest first		
tubulin a3f19c7e	Completed	14 files	33.8s	Today 14:22
Aspirin b71d0a44	Completed	15 files	61.2s	Today 13:05
EGFR c02be918	Running	queued		Today 12:48

The History tab listing past searches

Re-opening a past search shows its log. It does **not** restore its export files if those have already been deleted — re-run the search to get them back.

Part 7 — The AI analyst

The analyst answers questions about the search you are looking at. It receives the current query and search id as context.



The AI analyst panel with quick prompts and a conversation

Useful first questions:

- *Summarize the key findings of this search.*
- *Which compounds have the strongest binding evidence?*
- *Suggest follow-up queries worth running next.*

Two honest limitations:

- Answers are generated text, not a query over your files. Treat them as a starting point.
- Always verify a number you intend to cite against the exported CSV or workbook.

Analyst messages count against the plan allowance shown in your account panel.

Part 8 — Plans and payment

8.1 Buying a plan

1. Open your account panel and choose **Upgrade plan** (or **Pricing** if you are not signed in).

1. Each plan card shows the price, what it includes, and a **Pay ₹...** button.

The whole card is clickable.

1. You are taken to a secure Razorpay checkout that accepts cards, UPI and net banking.

1. After payment the server verifies the signature and activates the plan immediately. The panel confirms with your payment id.

If verification is slow, do not pay again. Reload the page — the server reconciles with the payment provider and will still credit the plan.

8.2 The Academic plan

The Academic plan is discounted for students and academic labs and requires verification.

1. Choose **Academic** on the plan grid. You are taken to the verification form.
2. Enter your institutional email address. Recognised academic domains are approved immediately.

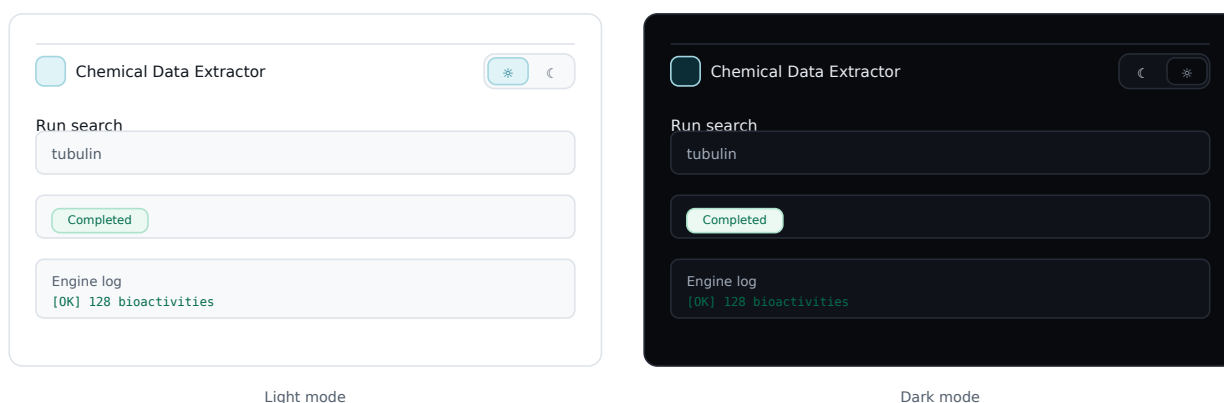
1. Choose your role and confirm your institution.
2. When the check passes, choose **Continue to payment**.

The verification is re-checked on the server when the payment is made, so an approval cannot be reused with a different account.

Part 9 — The interface

9.1 Light and dark mode

The theme switcher sits in the header, to the right of the section tabs.



The header with the theme switcher and the same page in both themes

Both themes are designed independently, not inverted: dark mode uses a near-black canvas with layered surfaces and light text, while light mode uses a white canvas with tinted surfaces and dark text. Your choice is stored in this browser and is restored the next time you open the application. If you have never chosen, the operating system preference is used.

9.2 Keyboard and screen readers

- Every control is reachable with **Tab** and has a visible focus ring.
- **Enter** submits the search form and the analyst question.
- **Escape** closes the account dialog.
- The engine log is a focusable, labelled region with aria-live updates while a search runs.
- Status is always carried by text as well as colour, so the interface is usable without colour perception.

9.3 Mobile and tablet

The workspace stacks into a single column below the tablet breakpoint: the form first, then the analyst, then results and the log. Section tabs move to their own row so they never crowd the theme switcher. Tables in the admin directory scroll horizontally instead of overflowing.

Part 10 — Troubleshooting

Message	Cause	Fix
The search engine wakes from idle...	The container was idle and is starting	Wait up to a minute and run again if needed
Waiting for another search	The single search slot is busy	Wait; your job is queued, not failed

Search lost — engine restarted	The server was replaced (deploy or free-tier recycle) while the job ran	Run the search again
Connection lost	The browser could not reach the engine six times in a row	Check your connection and run again
This search was lost because the server...	Same as above, reported earlier	Run the search again
Your search allowance is used up	The plan limit was reached	Upgrade the plan; the allowance resets next period
Create an account or sign in...	Authentication is required to run searches	Sign in from the header
Plans could not be loaded	The billing service did not answer	Close the dialog and try again
We could not verify this payment	The signature check failed	Do not pay again; reload and check the plan
The search produced very few results	The query name is not recognised	Try the full name, an alternative spelling, or a database identifier
A database step is amber in the log	That database timed out and was skipped	Nothing to do; other databases still contributed

The search takes longer than expected. Check the log before assuming failure. A 4-hop search legitimately runs for many minutes, and a status of *Running* with a moving bar means the engine is healthy.

Part 11 — Reference

11.1 Data sources

The search draws on 18 implemented data sources covering compounds, bioactivity, proteins, structures, pathways, genes, taxonomy, literature and patents. Every one is checked against its live API, and the interface lists their current status so you can see exactly what is available.

Area	Sources
Compounds	PubChem, ChEBI
Bioactivity	ChEMBL, PDBbind, BindingDB
Proteins	UniProt, STRING
Structures	PDBe-KB
Pathways	KEGG, Reactome, BRENDA, Gene Ontology
Genes and organisms	NCBI Gene, NCBI Taxonomy
Literature	Europe PMC, OpenAlex
Patents	PubChem patent cross-references and public patent records

Two things worth knowing, both shown in the interface rather than hidden:

- **Guide to Pharmacology** is unavailable on this deployment: the service now requires an API key that is not configured here.
- **BindingDB** holds *curated* affinities only. It answers protein accessions and compound names, but a compound with no curated measurement within the cutoff — some drug-like molecules simply have none — returns nothing. An empty BindingDB result means "nothing curated", not "nothing exists".

Seven of the sources answer only the identifier format they are authoritative for — a UniProt accession, a Reactome ID, a PDB code, a Gene Ontology term. If you search a plain name and they return nothing, search again with the accession for that entity and they will answer.

Patents

Patent records are reached through the compound cross-references held by PubChem, so coverage is strongest for well-known compounds — a search for *Aspirin* returns a large patent set, while an obscure compound may return none.

You can also search a publication number directly (US6280349, W02004018667 A1). The engine resolves the number to its title, assignee and filing year. That lookup depends on a public search endpoint which sometimes rate-limits; when it does, you get a direct patent search link for the number instead of a record, rather than an unrelated patent.

Patents appear in the graph as their own node type, linked to the compound they cover, and in a dedicated **Patents** sheet in the workbook.

11.2 Where results come from

A search builds its graph in six stages: expansion across the connected databases, entity resolution, deduplication, enrichment (filling in missing identifiers and properties), ranking by evidence, and export. The engine log shows each stage as it happens, and every node and relationship in the exports carries the database it came from, so any value can be traced back to its source.

- **Compounds and drugs** — PubChem, ChEBI, ChEMBL.
- **Proteins and structures** — UniProt, PDBe-KB, STRING.
- **Pathways and function** — KEGG, Reactome, BRENDA, Gene Ontology.
- **Genes and organisms** — NCBI Gene, NCBI Taxonomy.
- **Literature and patents** — Europe PMC, OpenAlex, and patent records reached through PubChem's compound cross-references.

11.3 Glossary

Term	Meaning
Node / entity	A thing discovered by the search
Edge / relation	A typed relationship between two nodes
Hop	One relationship step from the query
pChEMBL	Negative log of potency in molar units; higher is stronger
Enrichment	Filling in missing identifiers and properties after extraction
Export files	The files a finished search produces
Search id	The identifier assigned to a run, shown in the results header
Knowledge graph	The assembled set of nodes and their edges

11.4 Where to get help

- The **Resources** page of the application lists this tutorial, the quick reference and the export-format reference.
- The engine log of a failed run contains the most specific error available; quote the search id when reporting a problem.

Chemical Data Extractor v3.2 · multi-hop graph traversal · 18 verified data sources · enrichment pipeline