datagrok

An open-source, hackable, enterprise-ready chemical registration system

Oleksandra Serhiienko, Datagrok

https://github.com/datagrok-ai/mol-track

datagrok

Do we really need *another* registration system?

Yes, we do!

What's out there?

Commercial offerings:

- powerful, but expensive
- not self-hosted
- not easily customizable

Lwreg:

- great start!
- self-hosted and hackable

Who is left behind?


Academic labs focused on synthesis


Small biotechs just starting out


Organizations with legacy tools


Groups looking to explore new modalities

MolTrack: A hackable and flexible system

I have a dream…

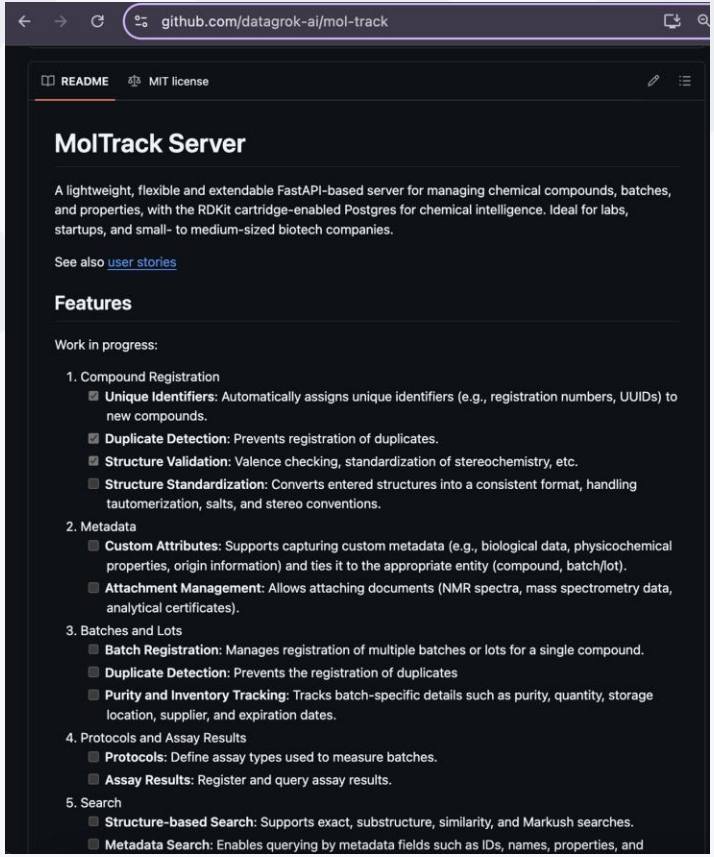
What if we could combine the ease and intelligence of commercial systems with the flexibility, affordability, and control of open-source?

MolTrack

- Fast
- Open-source
- Compound/batch/assay
- Hackable but enterprise-ready
- Dynamic properties
- Chemical intelligence (RDKit)
- Postgres-backed
- Searches, aggregations

Open-source

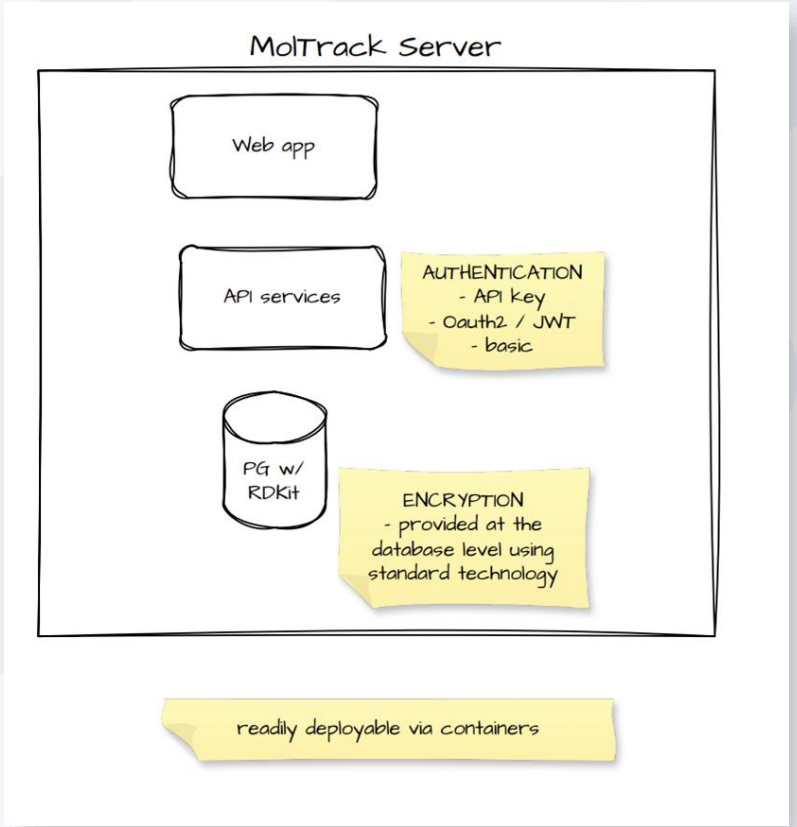
- MIT license
- Free as in beer
- Community can adapt the code to their own workflows



Postgres-backed

Leverages the power of relational databases:

- Transactions
- Backups and replication
- Querying and indexing
- Scalability
- Cloud RDS



Hackable

- MIT license
- Python plugin system
- Events and web hooks
- Modular architecture

…but enterprise-ready

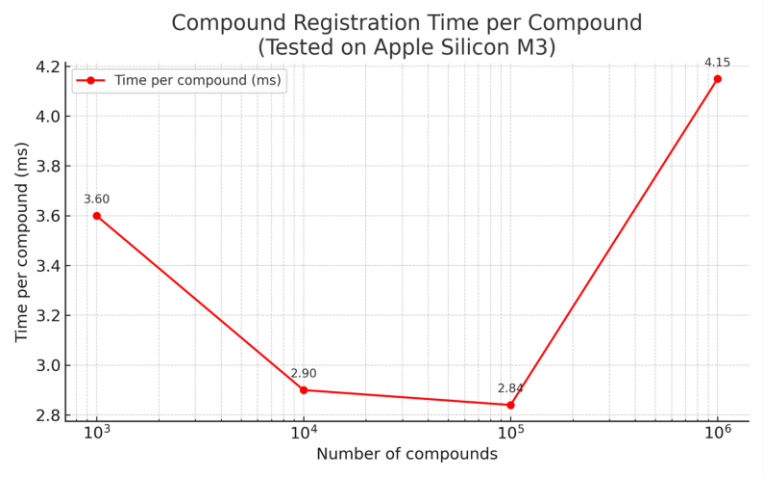
- Postgres: managed cloud databases, transactions…
- Authentication, API keys, tokens
- Users, roles, privileges
- Configuration
- Molecule-level and property-level permissions
- Audit trail
- Integration-ready

Get it up and running for your lab in a minute - or use as a backbone for the enterprise solution

Fast

ChEMBL

1000 compounds	3.57s	3.6 ms/compound
10,000 compounds	29.4s	2.9 ms/compound
100,000 compounds	284s	2.84 ms/compound
1,000,000 compounds	4149s	4.15 ms/compound

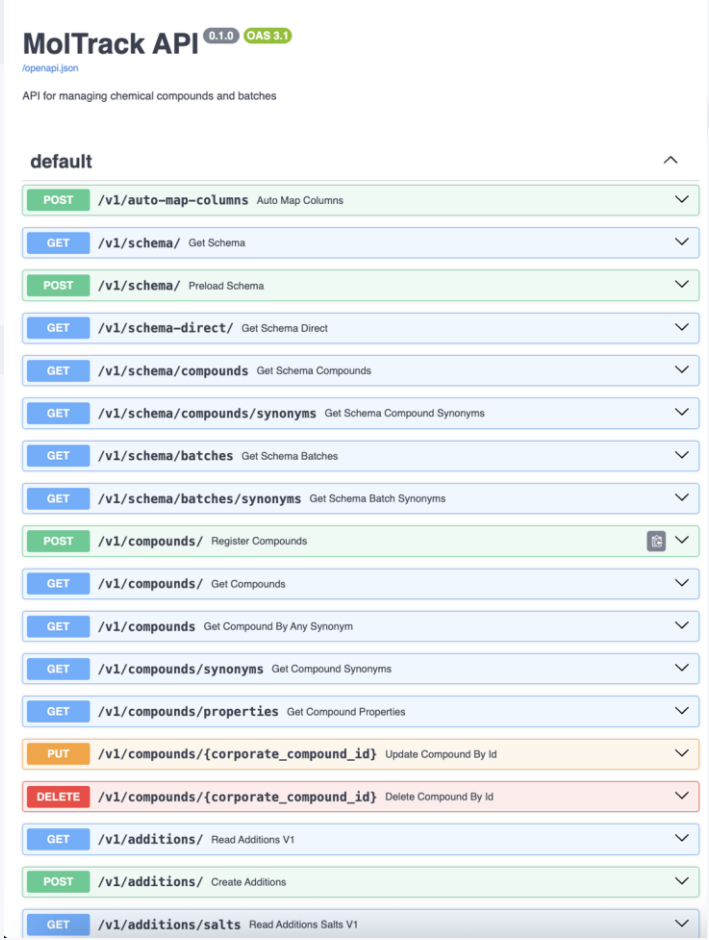


CLI

```
Usage: mtcli.py [OPTIONS] COMMAND [ARGS]...

 Options
 --install-completion          Install completion for the current shell.
 --show-completion             Show completion for the current shell, to copy it or
                               customize the installation.
 --help                        Show this message and exit.

 Commands
 schema       Schema management commands
 compounds    Compound management commands
 batches      Batch management commands
 properties   Property management commands
 additions    Addition management commands
 assays       Assays management commands
 database     Database management commands
 directory    Directory loading commands
 csv          CSV utility commands
 search       Search functionality
 admin        Administrative functions
```



CLI provides full access to Moltrack API capabilities or you can interact directly via the OpenAPI page.

Chemical Intelligence…powered by RDKit

- Cleanup
- Standardization
- De-duplication
- Validation
- Sketching
- Search via hash, substructure, similarity
- Configurable
- Extensible

- RDKit toolkit
- RDKit postgresql cartridge
- Standardization & Filtering methods
- Molecular Hashes

Administrators can define, extend, and enforce chemistry data quality

datagrok

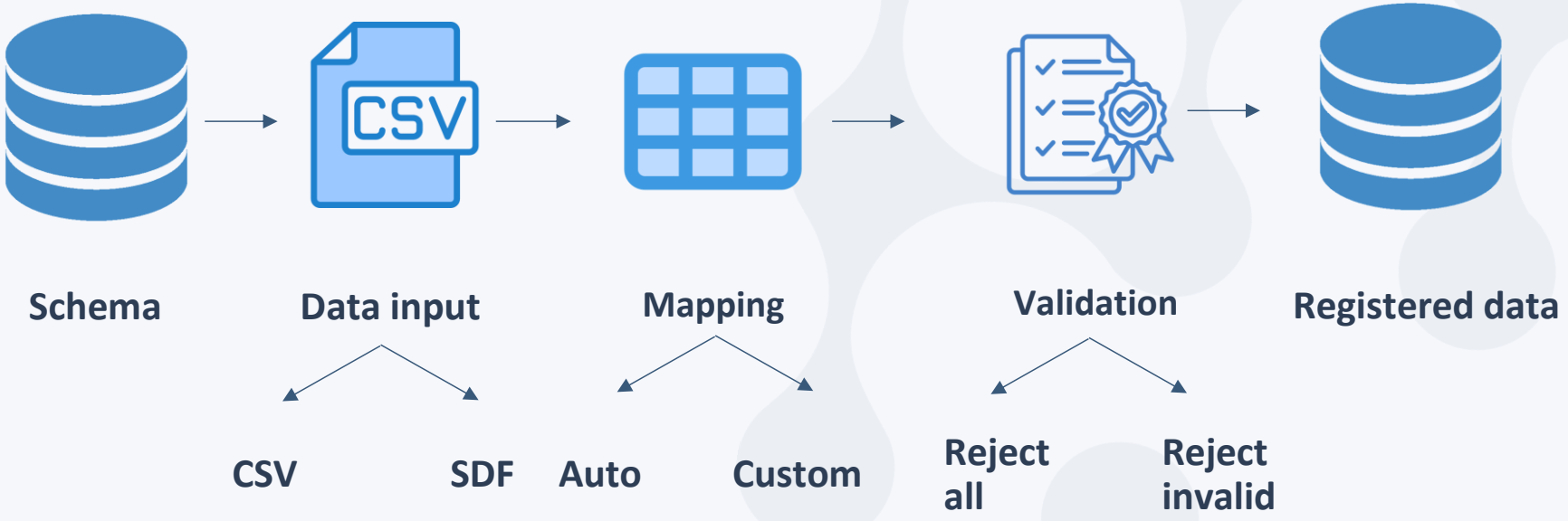
Schemas and properties

- Structured metadata for compounds, batches, assays
- Used for validation, field mapping
- Extensible with new properties

Properties

Name	Entity Type	Value Type
CAS	COMPOUND	string
CHEMBL ID	COMPOUND	string
Common Name	COMPOUND	string
corporate_compound_id	COMPOUND	string
IUPAC Name	COMPOUND	string
MolLogP	COMPOUND	double
Source	COMPOUND	string
Source Compound Code	COMPOUND	string
USAN	COMPOUND	string
Acquired Date	BATCH	datetime
corporate_batch_id	BATCH	string
ELN Reference	BATCH	string
EPA Batch ID	BATCH	string
Project	BATCH	string
Purity	BATCH	double
Responsible Party	BATCH	string
Source	BATCH	string
Source Batch Code	BATCH	string
Synthesized Date	BATCH	datetime
assay format	ASSAY	string
biological system	ASSAY	string
Assay Run Date	ASSAY_RUN	datetime
Assayer	ASSAY_RUN	string
Cell Concentration	ASSAY_RUN	double
Cell Lot	ASSAY_RUN	string
Cell Species	ASSAY_RUN	string
pIC50	ASSAY_RESULT	double

Registration:
Bringing data to life



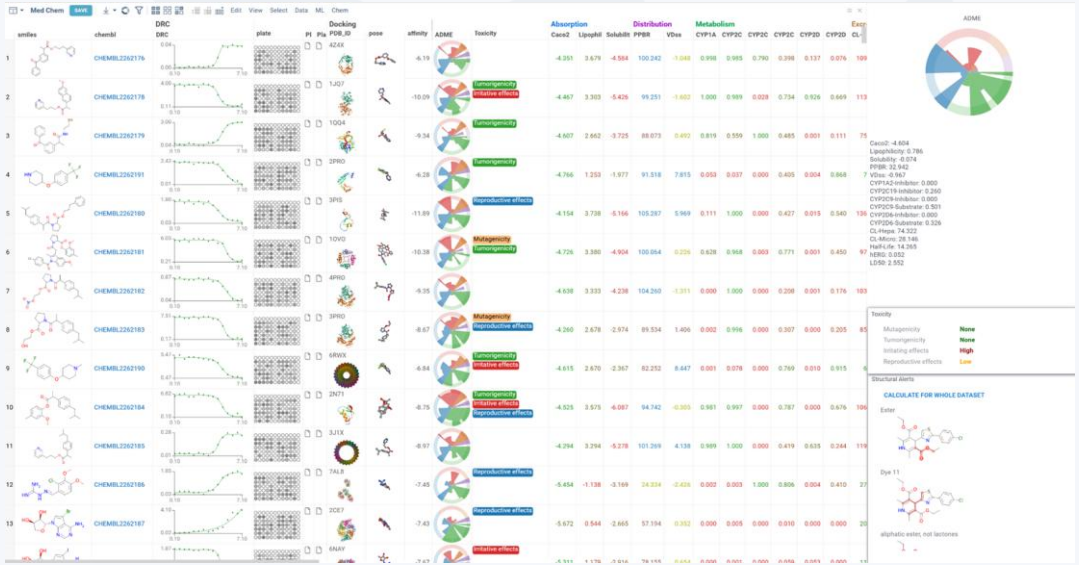
MolTrack UI: a Datagrok plugin

datagrok

Why Datagrok?

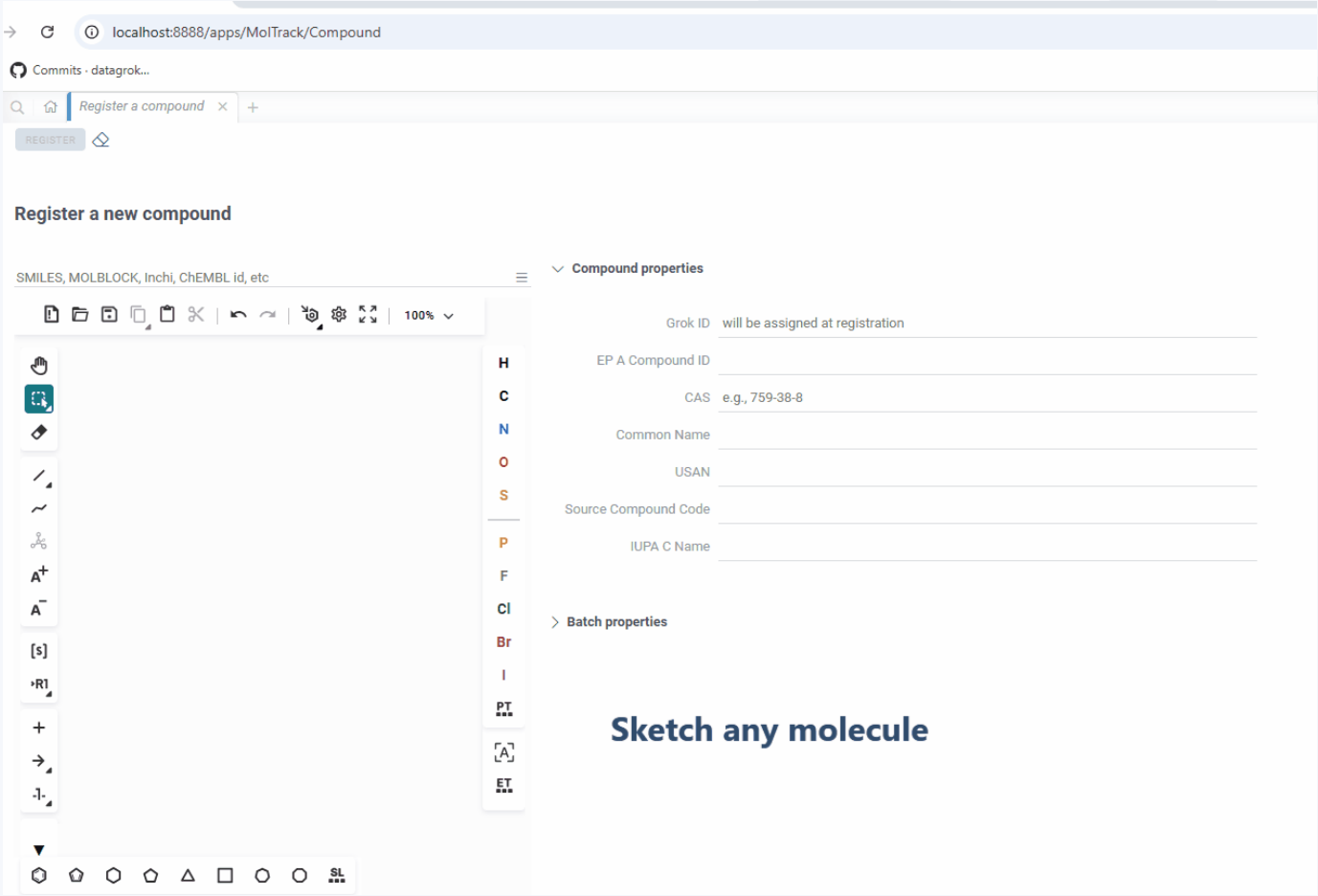
- Swiss-Army knife for cheminformatics
- First-class cheminformatics support
- Rendering, sketching
- Contextual info on molecules
- Connecting to databases
- ADME, Boltz-2
- Docking, folding, retrosynthesis
- Hit Design, Assay Plates

MolTrack service: MIT licensed
MolTrack plugin: MIT licensed; depends on Datagrok
Datagrok: proprietary; free for academic and personal use



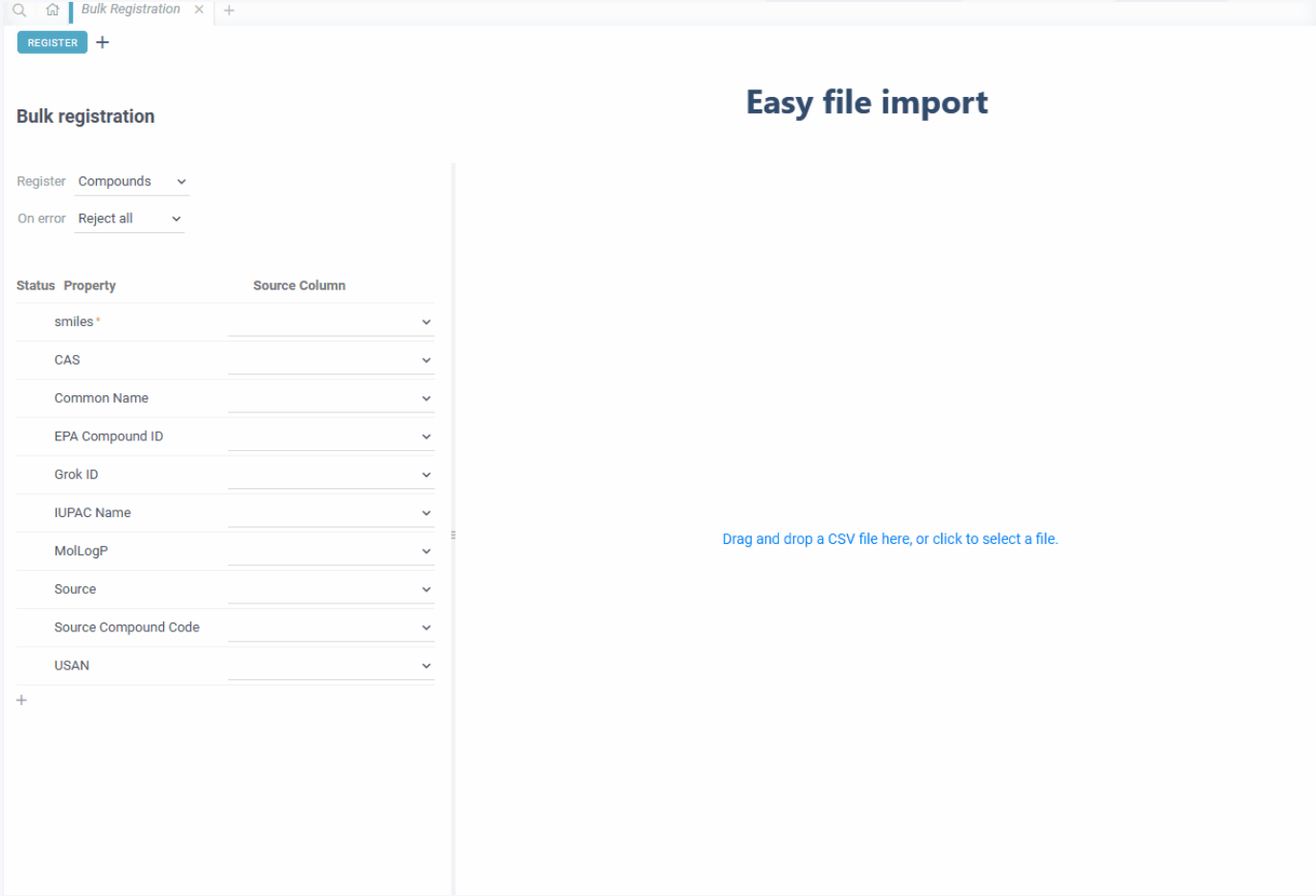
Single registration

datagrok



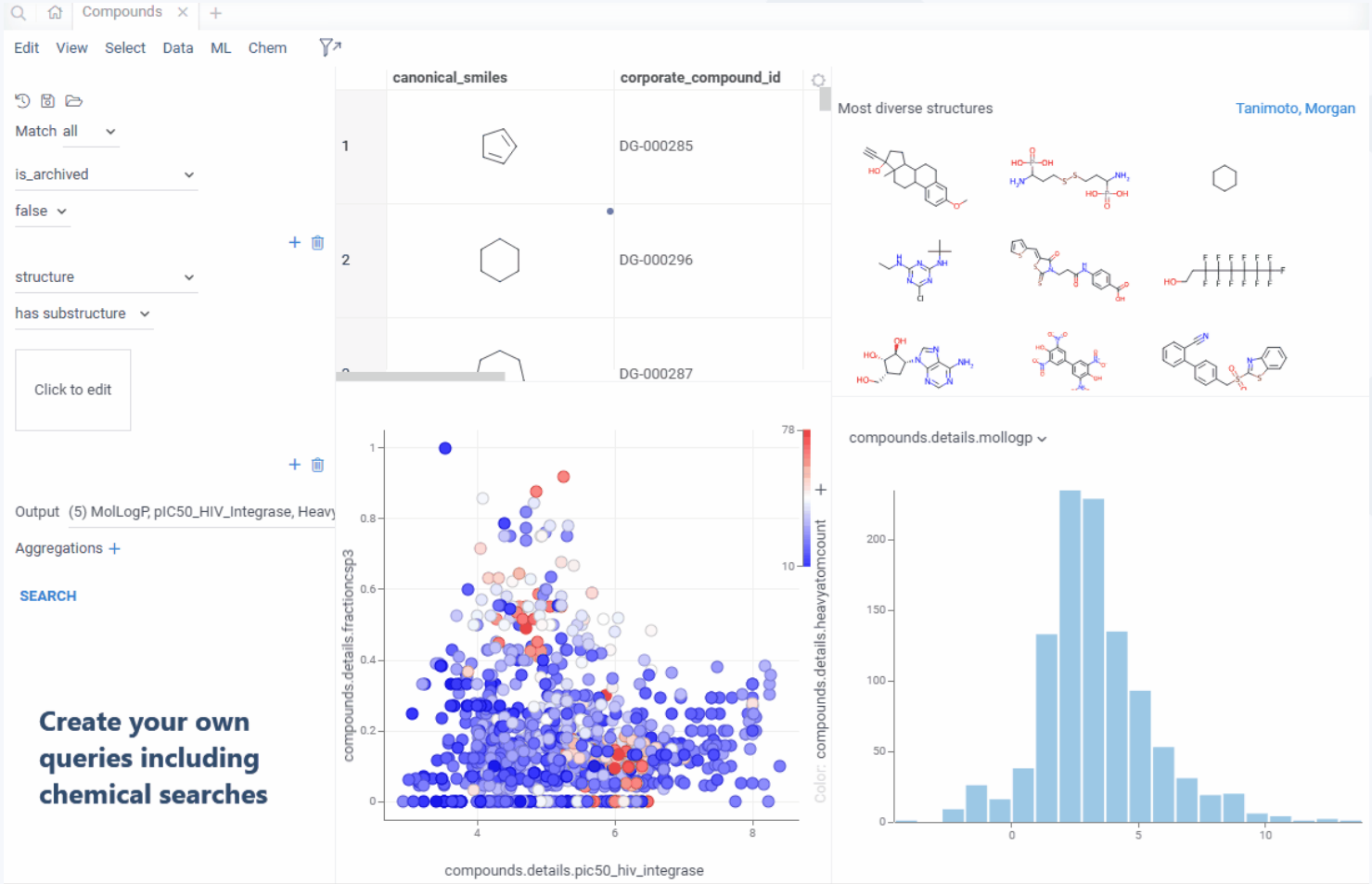
Bulk registration

datagrok



Search

datagrok



Create your own queries including chemical searches

Future plans

datagrok

- Additional modalities: no-structures, peptides, oligos, helm, biomolecules
- Synchronization with commercial registration systems
- Multiuser improvements (groups, roles, etc)
- Staging area to handle large files
- Monomer management
- Mixtures
- Computations and predictions

Acknowledgements

We want to thank RDKit community for the fantastic toolkit, and more specifically lwreg authors (Greg Landrum, Jessica Braun, Paul Katzberger, Marc Lehner, Sereina Riniker).

This has been a joint project between Datagrok and a company whose identity we cannot disclose yet. No further details at this time, but watch this space :)

Datagrok

- Oleksandra Serhiienko
- Ed Jaeger
- Maria Dolotova
- Andrew Skalkin

The TBA Company

- Danica Prodanovic Djapa
- Aleksei Khariukov
- Valentin Anoprenko

Thank You!