Categorical Data Integration for Computational Science:

→ What is Data Integration?

- The combination of two databases instances with different schemas into a coherent unified schema and instance
(How is combination done, mathematically?)
- Migration of data from one schema into another and querying a given database are specific cases of DI.

it can accelerate research

→ Problem being addressed:

There exist systemic issues in academic data sharing given the lack of a specific and universal standard for data representation (More so for heterogeneous data). The formalism of Category theory offers a flexible, machine interpretable language for declaring customized data structures and translating between them such that the intended meaning of the data provider and the data receivers are respected; mediating these transformations with CQL provides assurance of data quality by construction

[detectSystemType]
Note that merge & migrate are not primitive operations in CQL but rather they are the results of series of commands in CQL

→ Features of CQL:

1) • CQL schemas are not just simple data containers like SQL schemas, they also convey important conceptual knowledge
• They are structurally equivalent to ologs, which function like concept maps with a formal interpretation
• This means that:
→ Boxes in the schema can be seen as database tables
→ Arrows acts as foreign keys or functions
→ Equations enforce constraints that define relationship between entities.

2) • CQL enforces path equations, which act as integrity constraints that must be satisfied.
• These constraints prevent incorrect or nonsensical data from being stored in the database

3) Functorial Data migration, Hence structure-preserving.
Meaning, → Data from an external source can be used without violating the researcher's

constraints.

→ Data being shared can only be modified in ways that respect the original schema's structure and meaning.

4) • CQL provides automated & rule based transformations, eliminating many errors & inefficiencies that arise when manually adapting data to different schemas.

• CQL automatically prevents ~~such migration~~ invalid data migrations before they happen by analyzing schema mappings, rather than relying on post-migration validation

5) Since there is no single universal standard for data representation across disciplines, CQL provides flexible yet rigorous way to define custom schemas while maintaining standard structural integrity

This makes it possible to bridge different scientific data sets without sacrificing accuracy

Quantum chemistry datasets, especially those involving Density Functional Theory (DFT) calculations, present a challenge in standardization due to their complexity and variability. The lack of a universal format stems from several factors:

Diverse Software Ecosystems – Different DFT software packages (e.g., VASP, Gaussian, Quantum ESPRESSO) have their own input/output formats, making cross-compatibility difficult.

Arbitrary Preprocessing and Postprocessing – Many calculations involve custom workflows where input data is transformed before being fed into DFT software, and the outputs are further processed. This variability means that what constitutes a "calculation" can differ significantly between users and projects.

Multiple Structures and Configurations – A single "calculation" may involve multiple geometries, spin states, solvation models, or external perturbations, all of which contribute to complexity.

High Dimensionality of Data – DFT calculations generate not just energies but also wavefunctions, densities, forces, electronic properties, and other quantities, which may be stored in different formats or resolutions.

Metadata and Provenance Tracking – Because DFT calculations depend heavily on specific choices (e.g., exchange-correlation functionals, pseudopotentials, numerical grids), tracking and reproducing results requires detailed metadata that isn't standardized across the field.