

Contents

CHT.M1 – Molecular Data Acquisition and Representation	1
Project Overview	1
Research Question	2
Project Objectives	2
Project Workflow	3
Required Deliverables	3
Evaluation Criteria	3
Expected Outcome	4
Recommended Friendly Targets – Metabolic Disorders Context	4

CHT.M1 – Molecular Data Acquisition and Representation

Final Project. From Data Retrieval to a Structured Chemical Library

Version 1.0 - June, 2026. Monterrey

Author: [Flavio F. Contreras-Torres](#). Tecnológico de Monterrey.

Project Overview

This final project integrates the four notebooks developed throughout **Module 1** into a unified computational workflow. The objective is to transform raw molecular data retrieved from public databases into a **structured and visually interpretable chemical library** for a defined biological target.

You will assume the role of a Computational Chemist responsible for building a reproducible molecular dataset by combining **PubChem** and **ChEMBL** workflows. The emphasis is on technical execution, data acquisition, and molecular representation, ensuring that chemical records can be retrieved, connected, and interpreted before entering later stages of curation and analysis.

Project Estimation & Effort Breakdown

Project Phase	Estimated Time	Difficulty Level	Why does it take this long?
1. Selection & Extraction	2 - 3 hours	Medium	Selecting the biological target and navigating ChEMBL/PubChem IDs to ensure the API returns the desired data.

Project Phase	Estimated Time	Difficulty Level	Why does it take this long?
2. Cleaning & Standardization	3 - 4 hours	High	The most critical stage. Converting units, handling duplicates, and creating the activity labels often requires multiple code “refinements.”
3. Analysis & Visualization	2 - 3 hours	Medium	If previous notebooks are understood, this is mostly adapting existing plotting code to the new dataset.
4. Interpretation & Reporting	3 - 5 hours	High	This is where the student must think critically. Inspecting molecules and drafting SAR hypotheses requires careful observation.

Notice: This project is designed to be completed in approximately **10 - 15 hours**.

Student Checklist for Success

To ensure you complete the project within this timeframe, consider the following:

- **Start Simple:** Choose a well-studied target (e.g., Acetylcholinesterase, EGFR, or Mpro) to ensure plenty of data is available.
- **Reusable Code:** Leverage the functions and logic you built in Notebooks 1 through 4.
- **The 80/20 Rule:** Focus on getting a clean, functional dataset first; you can refine the visual styling of your plots later.

Research Question

How can raw molecular data retrieved from public databases be systematically transformed into a structured and visually interpretable chemical library for a specific biological target?

Project Objectives

1. **Programmatic Retrieval:** Extract structural data from PubChem using REST API queries.
2. **Advanced Structural Querying:** Implement at least one structure-based search using POST requests.
3. **Target-Centered Bioactivity Retrieval:** Retrieve ChEMBL bioactivity data for a selected biological target.

4. **Dataset Assembly:** Connect molecular identifiers, activity records, and biological context into a structured dataset.
5. **Molecular Representation:** Convert encoded molecular records into computational objects and 2D depictions.
6. **Structural Visualization:** Build a visual inspection panel of representative compounds from the final dataset.

Project Workflow

Phase 1 – PubChem Data Acquisition

Build an initial molecular dataset through REST API queries. Retrieve structural identifiers such as SMILES and InChIKey, along with selected computed molecular properties.

Phase 2 – Structure-Based Querying with POST Logic

Perform at least one **structure-driven** search using POST request. Demonstrate technical understanding of GET vs POST communication and document how chemical payloads differ from text-based queries.

Phase 3 – ChEMBL Bioactivity Retrieval

Select a biological target (e.g., PPAR γ , Mpro, GPCR, enzyme of interest) and retrieve experimental bioactivity records from ChEMBL. Organize activity values in nanomolar (nM), assay context, and molecular identifiers into a structured target-centered dataset.

Phase 4 – Molecular Representation and Visualization

Transform SMILES strings into molecular objects and generate 2D structural depictions for representative compounds. Use these visual outputs to inspect broad structural patterns and connect encoded records with chemical interpretation.

Required Deliverables

1. A final **structured dataset** in CSV format.
2. A set of representative 2D molecular visualizations from the selected target dataset.
3. A structured report in Markdown within the notebook, including:
 - selected target and biological rationale
 - summary of PubChem retrieval
 - example of POST-based structural querying
 - summary of ChEMBL bioactivity retrieval
 - explanation of how molecular and activity records were connected
 - brief visual interpretation of representative compounds
4. An enriched SDF file containing structural and associated metadata.

Evaluation Criteria

- **Reproducibility:** The workflow runs from raw retrieval to structured output without undocumented manual intervention.

- **Data Organization:** Molecular, biological, and activity records are connected clearly and consistently.
- **Technical Clarity:** Code is readable, documented, and aligned with the logic of the module.
- **Structural Interpretation:** The student demonstrates the ability to relate encoded molecular records to their visual chemical representations.

Expected Outcome

By the end of this project, you will have constructed a structured and visually interpretable chemical library derived through reproducible computational methods. This output will serve as the starting point for Module 2: Data Curation and Structural Exploration, where the dataset will be evaluated, standardized, and examined in greater analytical depth.

Recommended Friendly Targets – Metabolic Disorders Context

The following protein targets are strongly associated with obesity, insulin resistance, and metabolic syndrome. They are recommended due to their biological relevance, availability of bioactivity data in ChEMBL, and suitability for structure–activity analysis.

Target Protein	Biological Role in Metabolic Disorders	Pedagogical Justification
DPP-4 (Dipeptidyl Peptidase-4)	Regulates incretin hormones (GLP-1), directly linked to glucose homeostasis and type 2 diabetes.	Large and clean IC ₅₀ datasets in ChEMBL; well-characterized inhibitor scaffolds; ideal for binary classification and SAR visualization.
11β-HSD1 (11-beta- hydroxysteroid dehydrogenase type 1)	Controls intracellular glucocorticoid activation; associated with visceral obesity and metabolic syndrome.	Moderate-to-large dataset size; structurally diverse inhibitors; excellent for exploring lipophilicity–potency relationships.
AMPK (AMP-activated protein kinase)	Master regulator of cellular energy balance and metabolic adaptation.	Strong metabolic relevance; provides mechanistic depth; good diversity of activators and modulators for comparative profiling.
ACC1 (Acetyl-CoA Carboxylase 1)	Key enzyme in fatty acid biosynthesis; implicated in lipid accumulation and obesity.	Enzyme target with measurable IC ₅₀ values; useful for metabolic pathway discussion and structure–function reasoning.
FASN (Fatty Acid Synthase)	Catalyzes de novo lipogenesis; overactive in obesity and metabolic dysregulation.	Abundant inhibitor data; clear biochemical endpoint; suitable for potency standardization and physicochemical profiling exercises.

Recommendation

For maximum dataset stability and smoother workflow execution, **DPP-4** and **11 β -HSD1** are the most pedagogically robust alternatives.

License

The content of this tutorial itself is licensed under the terms and conditions of the [Creative Commons Attribution \(CC BY 4.0\) license](#), and the underlying source code used to format and display that content is licensed under the [MIT license](#). See the LICENSE files for full details.

Attribution

If you use or adapt this material, please provide appropriate credit to the original [authors](#) and repository: <https://github.com/NanoBiostructuresRG>