Accelerating Drug Discovery through Semantic Data Integration and Machine Learning: From Biomedical Knowledge Graphs to Predictive Models

Toshiaki Katayama[1*], Shuichi Kawashima[1], Yuki Moriya[1], Ryosuke Kojima[2,3], Takuto Koyama[3] and Mayumi Kamada[4]

[1] *Database Center for Life Science, Joint Support-Center for Data Science Research, Research Organization of Information and Systems, 178-4-4 Wakashiba, Kashiwa, Chiba 277-0871, Japan*

[2] *RIKEN Center for Biosystems Dynamics Research, Riken, Kobe, Hyogo, 650-0047, Japan*

[3] *Graduate School of Medicine, Kyoto University, 53 Kawahara-cho, Shogoin Sakyo-ku, Kyoto 606-8507, Japan*

[4] *School of Frontier Engineering, Kitasato University, 1-15-1 Kitasato, Minami-ku Sagamihara, Kanagawa, 252-0373, Japan*

Abstract

We introduce two interoperable resources that facilitate semantic data integration and machine learning in biomedical research: the Med2RDF knowledge graph and the Tabulae dataset preparation system. Med2RDF transforms heterogeneous biomedical databases into RDF using a unified ontology and publishes them via the RDF Portal with SPARQL endpoints and FAIR-compliant metadata. It addresses key integration issues such as identifier heterogeneity and vocabulary inconsistency.

Tabulae builds on this semantic infrastructure by enabling users to efficiently generate machine learning–ready tabular datasets from complex, RDF-based sources. It abstracts the complexities of querying and integrating diverse information, including compound and protein features, into a unified and readily usable format for machine learning algorithms.

We demonstrate the utility of these resources through a case study on compound–protein interaction (CPI) prediction using Random Forest regression. This illustrates how semantic integration combined with machine learning can support efficient drug discovery, and highlights Tabulae's unique capability in bridging the gap between rich knowledge graphs and the data formats required by modern machine learning workflows. Together, Med2RDF and Tabulae provide a scalable and reusable framework for semantic data-driven research in biomedicine.

Keywords

Knowledge graph, RDF, SPARQL, Machine learning, Med2RDF, Tabulae

1. Introduction

The landscape of biomedical research is characterized by an ever-growing volume of diverse and heterogeneous datasets, ranging from chemical compounds and protein sequences to

ISWC 2025 Companion Volume, November 2-6, 2025, Nara, Japan

* Corresponding author.

ktym@dbcls.jp (T. Katayama); kwsm@dbcls.rois.ac.jp (S. Kawashima); moriya@dbcls.rois.ac.jp (Y. Moriya); ryosuke.kojima@riken.jp (R. Kojima); koyama.takuto.82j@st.kyoto-u.ac.jp (T. Koyama); kamada.mayumi@kitasato-u.ac.jp (M. Kamada)

0000-0003-2391-0384 (T. Katayama); 0000-0001-7883-3756 (S. Kawashima); 0000-0001-8195-5893 (Y. Moriya); 0000-0003-1095-8864 (R. Kojima); 0000-0002-9569-8370 (T. Koyama); 0000-0002-2555-7345 (M. Kamada)

clinical observations and scientific literature. While these resources contain highly complementary information, their fragmentation and semantic inconsistencies pose significant challenges to comprehensive data analysis, hindering advanced applications like drug discovery and clinical decision support. Traditional approaches to data integration often fall short due to disparities in identifier schemes, property vocabularies, and conceptual modeling. To address these fundamental challenges, the Semantic Web technologies, particularly the Resource Description Framework (RDF), offer a powerful paradigm for expressing and linking such disparate data into a unified, machine-readable format, thereby enabling sophisticated cross-database queries and inference [1, 2]. Furthermore, the integration of knowledge graphs with machine learning has emerged as a critical area, offering potential for enhanced interpretability and more robust model building in complex domains like biomedicine [3, 4]. This paper details a comprehensive approach to biomedical data integration and its application in machine learning for drug discovery, highlighting the foundational work of the Med2RDF project and RDF Portal [5], and introducing Tabulae as a novel system for preparing integrated datasets for machine learning workflows.

2. Med2RDF Project and RDF Portal: Building a Foundation for Semantic Integration

The Med2RDF project is a significant initiative dedicated to constructing a semantically integrated biomedical knowledge graph by converting diverse life science databases and documents into RDF. Initial work on Med2RDF has been previously presented, detailing its conceptual framework and early applications [6]. This project directly confronts the challenges of identifier heterogeneity, property inconsistency, and class misalignment through the development of a dedicated Med2RDF ontology and an ecosystem of interoperable tools. The RDF conversion pipeline involves data ingestion from original sources, mapping to RDF using custom converters, semantic alignment via the Med2RDF ontology, and subsequent publication through the RDF Portal and SPARQL endpoints. The Med2RDF ontology is crucial for providing unified URIs and classes, harmonizing predicates through reusable properties, and aligning with established external ontologies such as MeSH, ChEBI, and OBO Foundry, thus enabling consistent annotation and facilitating cross-source reasoning and federated querying. Currently, the Med2RDF project has successfully RDFized a wide array of databases, including variation databases (dbSNP, dbVar, dbNSFP, gnomAD, ClinVar), cancer/clinical data (COSMIC, CIViC), drug/target information (DGIdb), and protein/gene resources (HiNT, INstruct, HGNC). It also integrates existing RDF datasets like UniProt, PubChem, ChEMBL, Rhea, ChEBI, MeSH, PDB, and DDBJ via the Med2RDF ontology, fostering a highly interconnected data environment. The RDF Portal (https://rdfportal.org) serves as the central public platform for hosting these RDF datasets, providing SPARQL endpoints, VoID metadata, and comprehensive dataset catalogs, all while promoting FAIR (Findable, Accessible, Interoperable, Reusable) principles. Furthermore, TogoID [7], a complementary system, enhances this integration by mapping relationships between database identifiers using a shared ontology, thereby enabling uniform expression of cross-links even for non-RDF datasets and significantly contributing to a unified, richly connected life science knowledge graph.

3. Tabulae: A Support System for Generating Integrated Datasets for Machine Learning

Building upon these foundational efforts in semantic data integration, we have developed Tabulae (https://github.com/dbcls/tabulae), a novel system designed to facilitate the generation of integrated datasets from various underlying databases for machine learning applications (Figure 1). Tabulae aims to bridge the gap between complex, semantically integrated biomedical data and the tabular formats often required by machine learning algorithms. While the underlying mechanisms may involve sophisticated data acquisition techniques, including querying multiple databases via SPARQL endpoints or leveraging pre-processed knowledge graphs, Tabulae provides a streamlined approach for users to obtain readily usable datasets. This system aggregates crucial information, such as compound properties (e.g., SMILES, Ro5, AlogP, MW, HBA, HBD) and protein details (e.g., UniProt IDs, sequences) from sources like chembl_protein table. By abstracting the complexities of data integration, Tabulae enables researchers to quickly access and prepare high-quality, pre-integrated data suitable for immediate use in computational experiments, significantly accelerating the initial data preparation phase in various biomedical applications, including drug discovery.

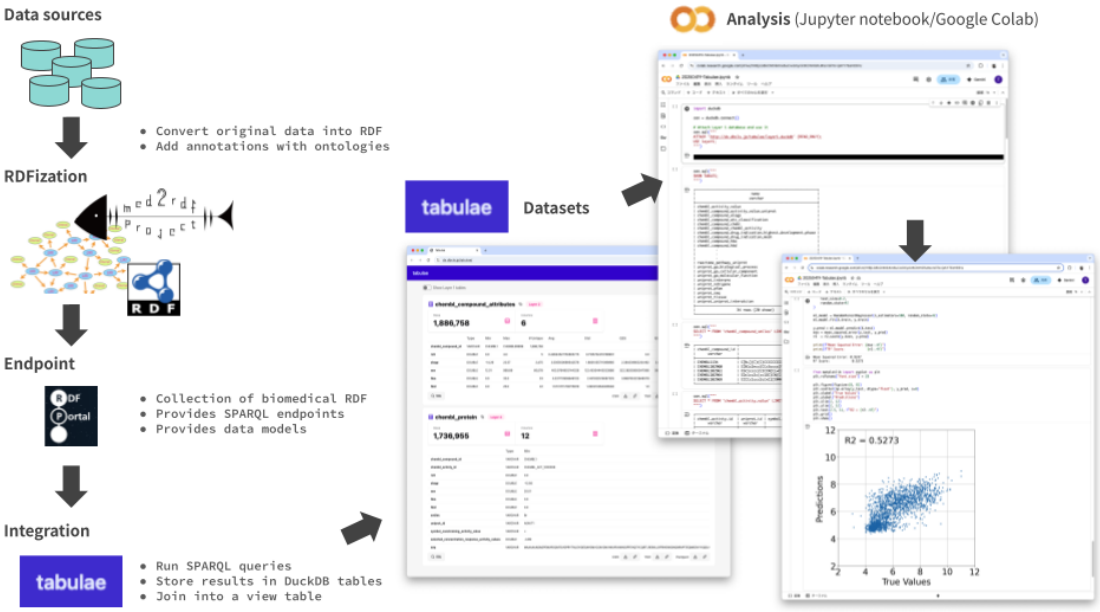

Figure 1: Flow of Data Standardization by Med2RDF and Data Integration and Analysis by Tabulae

4. Case Study of CPI Prediction: A Practical Application of Tabulae

To demonstrate the utility of Tabulae in practical drug discovery workflows, we present a case study focused on Compound-Protein Interaction (CPI) prediction using a machine learning-based approach. For this analysis, a subset of approximately 10,000 data points with specified activity values ("=") was randomly sampled from the Tabulae-acquired chembl_protein table. Given the computational demands of protein language models, the analysis was executed within a T4 GPU runtime environment. The methodology involved a comprehensive feature

engineering pipeline to transform raw chemical and biological data into numerical representations suitable for machine learning. Compound features were generated from SMILES strings using Morgan Fingerprints, resulting in a feature dimension of 512. SMILES (Simplified Molecular Input Line Entry System) is a textual representation of chemical structures, while Morgan Fingerprints are widely used chemical descriptors that encode structural information of a molecule into a fixed-length binary vector. For protein features, amino acid sequences were converted into embeddings using the ESM2 protein language model (facebook/esm2_t6_8M_UR50D), yielding a feature dimension of 320. ESM2 is a state-of-the-art protein language model developed by Meta AI, capable of generating rich contextual embeddings from protein sequences, capturing complex functional and structural information. These distinct compound and protein feature vectors were then concatenated to form a unified input feature set, resulting in a total feature dimension of 832. This combined feature set was then utilized to construct a regression model aimed at predicting selected concentration response activity values.

A Random Forest Regressor with 100 estimators was employed for model construction. The dataset was split into 80% for training and 20% for testing to ensure robust evaluation of the model's generalization capabilities. The constructed model was evaluated on the unseen 20% test set, with its predictive performance assessed using widely accepted metrics: Mean Squared Error (MSE) and R-squared (R2) score. Furthermore, the predictive accuracy was visually confirmed through a scatter plot comparing true and predicted activity values, effectively illustrating the model's ability to capture and predict complex CPI relationships. This case study not only highlights the practical utility of Tabulae as a valuable resource for drug discovery research by providing easily accessible, integrated data, but also exemplifies a robust machine learning workflow for predicting CPIs. The successful demonstration of this approach underscores its potential to accelerate the identification of novel drug candidates and to streamline the early stages of the drug discovery process.

5. Conclusion

In conclusion, the Med2RDF project and the RDF Portal have established a robust framework for the semantic integration and dissemination of heterogeneous life science data, laying the groundwork for a comprehensive biomedical knowledge graph. Complementing these efforts, Tabulae provides a crucial link by enabling the efficient extraction and preparation of integrated, machine learning-ready datasets from these rich data resources. The successful application of Tabulae-derived data in the Compound-Protein Interaction prediction task validates its utility in real-world scenarios, showcasing its power in generating insights for drug discovery. We anticipate that this integrated approach, from knowledge graph construction to data preparation for machine learning, will be broadly applicable to numerous life science and medical applications, including target identification, drug repurposing, personalized medicine, and systems biology research, ultimately advancing data-driven decision-making across the biomedical domain.

Acknowledgements

The authors thank Dr. Yoji Shidara of Enishi Tech Inc. for his contributions to the implementation of Tabulae.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT and NotebookLM in order to: Text translation, Grammar and spelling check. After using these services, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] Erick Antezana, Martin Kuiper, and Vladimir Mironov. "Biological knowledge management: the emerging role of the Semantic Web technologies" Brief in Bioinformatics 10(4):392-407 (2009). doi: 10.1093/bib/bbp024.

[2] Ilaria Tiddi and Stefan Schlobach. "Knowledge graphs as tools for explainable machine learning: A survey" Artificial Intelligence 302, 103627 (2022).

[3] Jason Youn, Navneet Rai, and Ilias Tagkopoulos. "Knowledge integration and decision support for accelerated discovery of antibiotic resistance genes" Nature Communications 13(1):2360 (2022). doi: 10.1038/s41467-022-29993-z

[4] Yashrajsinh Chudasama, Hao Huang, Disha Purohit, and Maria-Esther Vidal. "Toward Interpretable Hybrid AI: Integrating Knowledge Graphs and Symbolic Reasoning in Medicine" IEEE Access 13, pp. 39489-39509 (2025), doi: 10.1109/ACCESS.2025.3529133.

[5] Shuichi Kawashima, Toshiaki Katayama, Hideki Hatanaka, Tatsuya Kushida and Toshihisa Takagi. "NBDC RDF portal: a comprehensive repository for semantic data in life sciences" Database (Oxford). bay123 (2018). doi: 10.1093/database/bay123.

[6] Mayumi Kamada, Toshiaki Katayama, Shuichi Kawashima, Ryosuke Kojima, Masahiko Nakatsui, and Yasushi Okuno. "Med2RDF: Semantic Biomedical Knowledge-base and APIs for the Clinical Genome Medicine". SWAT4HCLS 2019, pp. 161-162 (2019).

[7] Shuya Ikeda, Hiromasa Ono, Tazro Ohta, Hirokazu Chiba, Yuki Naito, Yuki Moriya, Shuichi Kawashima, Yasunori Yamamoto, Shinobu Okamoto, Susumu Goto, Toshiaki Katayama. "TogoID: an exploratory ID converter to bridge biological datasets" Bioinformatics; 38:4194-4199 (2022). doi: 10.1093/bioinformatics/btac491.