

Semantic Materials Science for Large Scale Molecular Representation Learning

Quynh D. Tran^{1,2}, Meredith A. Francis^{1,2}, Isabella M. Giammattei^{1,2}, Holly L. Schreiber^{1,3}, Katherine Q. Karkosiak^{1,2}, Erika I. Barcelos^{1,2}, Ozan Dernek^{1,2}, Laura S. Bruckman^{1,2}, Yinghui Wu^{1,3} and Roger H. French^{1,2,3,*}

¹The Center for Materials Data Science for Stockpile Stewardship, Case Western Reserve University, 10900 Euclid Ave Cleveland Ohio 44106

²Materials Science and Engineering Department, Case Western Reserve University, 10900 Euclid Ave Cleveland Ohio 44106

³Computer and Data Sciences Department, Case Western Reserve University, 10900 Euclid Ave Cleveland Ohio 44106

Abstract

Representation learning of molecules for molecular and materials discovery is a challenging task given the complex and multiple hierarchical levels that contribute to property and structural relationships. There are multiple naming schemas and ways to represent molecules that add to the challenge of applying machine learning (ML) to the field of molecular modeling. Artificial intelligence (AI) and ML techniques are useful for finding hidden complex relationships in non-linear systems with high dimensions. The implementation of semantic data management (SDM) for machine understanding of complex systems of molecules can guide AI/ML towards more accurate reasoning and learning using ontology and Resource Description Framework (RDF) linked graphs (i.e. knowledge graphs). In this paper, we utilize semantic materials science to develop an expansive energetic materials (EM) ontology, adapt historical molecular data into FAIR (findable, accessible, interoperable, reusable) data, and build a large, AI-ready dataset of over 4 million existing molecules from open-source chemical databases.

Keywords

Molecular Discovery, Semantic Data Management, Semantic Materials Science, Representation Learning

1. Introduction

Artificial intelligence (AI) and machine learning (ML) are critical tools in chemistry and materials for learning and extracting insights across complex datasets in non-linear heterogeneous systems. Examples of applications of AI/ML to expedite chemistry and materials science research include property prediction and discovery of novel materials and molecules with predicted functions. Massive, unbiased, and curated datasets are needed to unleash and harness the predictive power of AI/ML models with high accuracy and trustworthiness. Materials science and chemistry experiments are slow and complex and high quality annotated datasets are expensive to generate unlike images and texts in computer vision and large language models. The lack of standards for recording metadata and missing critical experimental details mean integration of historical data is difficult and frequently data is un-interoperable and discarded. Limited data reuse severely constrains AI/ML applications in chemistry and materials science domains to expedite research [1].

The best models can be inaccurate and misleading if trained on biased or inaccurate data, thus, dataset bias is another critical challenge in the field of AI/ML. The existing molecular dataset is limited by current biases based on chemical intuition and conventions of the field. Small or filtered datasets

SeMatS 2026: The 3rd International Workshop on Semantic Materials Science co-located with the 25th International Semantic Web Conference (ISWC 2026), October 25th, 2026, Bari, Italy.

*Corresponding author.

✉ qdt@case.edu (Q. D. Tran); maf203@case.edu (M. A. Francis); img24@case.edu (I. M. Giammattei); hls92@case.edu (H. L. Schreiber); kqk2@case.edu (K. Q. Karkosiak); eib14@case.edu (E. I. Barcelos); oxd37@case.edu (O. Dernek); lsh41@case.edu (L. S. Bruckman); yxw1650@case.edu (Y. Wu); rxf131@case.edu (R. H. French)

ORCID: 0009-0006-4025-1834 (Q. D. Tran); 0009-0006-1325-7018 (M. A. Francis); 0009-0009-7102-2796 (I. M. Giammattei); 0009-0008-8229-5776 (H. L. Schreiber); 0000-0001-7951-4781 (K. Q. Karkosiak); 0000-0002-9273-8488 (E. I. Barcelos); 0000-0002-5071-9684 (O. Dernek); 0000-0003-1271-1072 (L. S. Bruckman); 0000-0003-3991-5155 (Y. Wu); 0000-0002-6162-0532 (R. H. French)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

can introduce unknown biases in the training data. Additionally, un-interoperable historical data also adds to potential unknown biases. Besides data limitations, there are several unique challenges associated with representing molecular data as a whole, including inconsistent molecular naming and identification conventions and the complex process of developing useful and informative molecular representations for downstream ML tasks. Successful representation of compounds for ML requires the use of unique and accurate chemical identifiers, which is a persistent challenge in chemistry. Many naming conventions have been developed to systematically classify and describe molecular structures, but they can be ambiguous, non-machine-readable or limited in the amount of structural information they can convey.

While large molecular datasets such as PubChem, ChEMBL, Cambridge Structural Database (CSD), and USPTO patent literature exist, there is no existing work that combines all of the molecules for molecular prediction and discovery [2, 3, 4, 5]. Existing machine learning and deep learning studies focus on benchmark and computed datasets such as ZINC20 and GDB-13 [6, 7, 8, 9, 10]. In niche domains such as energetic materials (EM) research, even less data is available for training generalizable models for predicting detonation and sensitivity parameters of energetic molecules and generating novel energetic molecules. Besides the inherent danger of handling and testing energetic materials, their performance and insensitivity are highly dependent on processing and experimental conditions which are commonly omitted. Thus, experimental results with missing/incomplete metadata are commonly contradictory and cannot be integrated to feed into deep learning workflows.

To alleviate dataset biases and small dataset challenge in task-specific molecular property prediction and discovery, we propose a framework for curating and integrating massive molecular datasets for representation learning using semantic materials science. Our contribution in this paper includes: 1) expanding the existing MDS-Onto ontology to accommodate for molecular data, 2) curating and integrating large open-source databases using MDS-Onto, and 3) demonstrating an example of representation learning with a curated large molecular dataset. The integrated dataset will enable researchers to answer wide array of scientific questions and benchmark and validate deep learning models using different subsets of data for training, validation, and testing.

Representation learning is a set of techniques or strategies used in ML for chemistry that automates learning by optimizing mappings in the embedding space to reflect structures or relationships in real data [11]. Representation learning can broaden molecular discoveries by developing an unbiased chemical latent space, but building a comprehensive chemical latent space requires the integration of larger, comprehensive datasets. In this paper, we apply molecular representation learning to EM with a different approach. We created a comprehensive chemical latent space of existing molecules from various databases by applying semantic materials science to the field of chemistry for the integration of large scale historical data instead of filtering our dataset to focus on areas of concentrated energetic molecules.

Semantic materials science is a strategy to build and curate AI-ready chemistry data for molecular property prediction and molecular discovery. Semantic data management (SDM) components include ontologies, knowledge graphs, study protocols, and FAIR (findable, accessible, interoperable, reusable) principles to reduce variability, ensure reproducibility, and allow data reuse and interoperability. FAIR guiding principles ensure scientific data is properly recorded and stored for transparency [12, 13]. Implementing SDM in energetic materials data allows for the reuse and reformatting of existing historical data into AI-ready data for AI/ML analysis by maintaining metadata, and providing a knowledge backbone, critical for scientific discovery [14]. While data specifically on EM are limited, SDM allows for the integration of large chemical databases, to provide knowledge on chemical structure, property and function, for a comprehensive learning pool. We apply SDM to the field of energetic materials to first develop an ontological domain for energetic materials in the MDS-Onto ontology for materials science. We then use this ontology to transform historical data into FAIR, ontology-linked data ready for AI/ML analysis, and integrate large, available databases of chemical structure, property, and reaction data. Lastly, we demonstrate the application of these methods on the organization and exploratory data analysis of 3.7 million molecules. Our work demonstrates generalizable datasets, tools, and methods of semantic materials science successfully applied to chemical and energetic material applications.

2. Molecular Identifiers & Representations

Many naming conventions have been developed to systematically classify and describe molecular structures, but they can be ambiguous and face trade-offs between the amount of structural information conveyed and readability. Traditional naming conventions, such as the International Union of Pure and Applied Chemistry (IUPAC) naming system, provide enough information for chemical structures to be deduced by human chemistry experts, however, allow multiple names to exist for the same molecule [15]. IUPAC compound names also do not store structural information that machines and models use for learning and reasoning. Chemical or molecular formulas, which represent molecules as a string, are machine-readable and compact (e.g.: glucose is represented as $C_6H_{12}O_6$); however, they do not contain atomic connectivity or 3D structural information.

In 1987, the Simplified Molecular Input Line Entry System (SMILES) was proposed to represent molecules and compounds as strings for computer understanding [16]. ReactionSMILES was developed as a notation for all of the molecular components in a reaction: reactants, reagents, solvents, products, catalysts, etc. SMILES strings can be generated by a number of different computational methods; thus, multiple valid SMILES can represent the same molecule. For example, COO and OOC are both valid SMILES strings that represent the same underlying carbon dioxide molecule. Stereochemistry can also be dropped from SMILES strings during parsing, e.g. C/C=C/C (E) or C/C=C\C (Z) is parsed into CC=C(C) or CCCC, then one SMILES string can represent multiple spatially distinct molecules. ChemAxon Extended SMILES (CXSMILES) and ReactionCXSMILES were developed by ChemAxon as a method to capture richer structural information and complement SMILES strings [17].

IUPAC Chemical International Identifiers (InChI) was developed in 2005 and are built to uniquely identify molecules, unlike SMILES, however, they are not human-readable [15, 18]. InChI Keys are generated from a hash function into a 27-digit string from the InChI string. SMILES stores structural information about the molecules, while InChI, IUPAC name, and InChIKey do not. This adds another layer to uniquely identifying and integrating large molecular databases.

In addition to correctly identifying molecules, a distinct challenge in ML for chemistry is accurately representing molecular data for model learning and encoding critical information. Molecular representations are simplified models of molecules containing structural and property information of the real molecules that can be understood by a computer. This is a non-trivial task, as molecular functionality depends on non-linear, complex interactions between atoms, bonds, functional groups, and environmental factors such as other molecules. There is a vast array of computational methods to represent 3D molecules developed over the years. The most common representations are:

- **Chemical descriptors:** a set of properties and structural information used to describe molecules. Examples include molecular mass, dipole moment, vibrational zero-point energy, HOMO-LUMO gap, and specific heat.
- **One-dimensional:** representative sequence strings, examples including Simplified Molecular Input Line Entry System (SMILES), IUPAC International Chemical Identifier (InChI), Self-Referencing Embedded Strings (SELFIES), and Extended-Connectivity Fingerprints (ECFP) [16, 18, 15, 19, 20].
- **Two-dimensional:** molecules are represented as graph models, which model relationships between nodes. In molecular graphs, atoms are indicated as nodes and bonds as edges.
- **Three-dimensional:** molecular representations with 3D properties remains a grand challenge because 3D structural information is difficult and expensive to generate and collect. Thus, there is limited 3D data for learning and training [21].

Chemical descriptors are commonly used in statistical ML models (e.g. linear regression, logistic regression, K-nearest neighbors, random forest, etc.) for Quantitative Structure-Property Relationship (QSPR) [22]. The set of chemical descriptors used for predictive modeling is chosen based on their high correlation to the properties of interest. One-dimensional molecular representations do not store proper structural information such as atomic connectivity, chirality, bond length, or bond angle. Extended-Connectivity Fingerprints (ECFP), also called Morgan Fingerprints, contains neighborhood structural

information as bit vectors (1024 or 2048). They can also be ambiguous and non-unique (i.e. SMILES). In recent years, novel systems such as Self-Referencing Embedded Strings (SELFIES) and SMILES Arbitrary Target Specification (SMARTS) are developed to increase robustness and expressivity while maintaining chemical validity. SELFIES ensure valid representation of molecules where every combination of symbols from its alphabet maps to a chemically valid graph. However, they are longer than SMILES strings to ensure robustness [23]. SMARTS are patterns used to identify specific molecular substructures. They allow to search specific molecular properties such as functional group, atoms or bonding rules but fall short on readability and struggle to describe 3D information [20]. Two-dimensional representations (i.e. molecular graphs) include proper atomic connectivity, atom features, bond features, etc.; however, are missing 3D structural information that influences properties such as solubility and density and is critical for accurate models. SMILES and molecular graphs are the most common representations in the literature but lack the 3D information that is critical for predicting molecular properties. The challenge of establishing non-linear, complex structure-property relationships remains.

3. Chemical Databases

There are a number of different databases for open-source and proprietary purposes for large datasets of molecules and their properties which often use their own system to track molecules. Examples of these open-source databases include the PubChem database which contains chemical name, molecular formula, structure, and other chemical and physical properties. Molecular data from PubChem can be obtained via the pubchempy package (Python API) or via Resource Description Framework (RDF)-linked graphs, which only contains calculated physicochemical properties [2]. The Cambridge Structural Database (CSD) contains crystallographic data in its majority and can be extracted using the Cambridge Crystallographic Data Centre (CCDC) Python API. This database is highly curated as submissions are reviewed by in-house experts before publishing. This API provides high-throughput access to 3D atomic coordinates and other experimental unit cell data [4]. The ChEMBL database contains chemical, bioactivity and genomic data for drug discovery. Its data was extracted using its own repository designed for easy access and reuse [24, 3, 25]. The Open Reaction Database (ORD) data consists of 2.4 million reactions and organic reaction data submitted by various researchers and labs. It was accessed by the Open Reaction Database's ord-data Github repository [5]. The Chemical Entities with Biological Interest (ChEBI) is an ontology for chemical compounds in biological context; connected to BFO [26]. There are a number of subscription-based databases of chemical compounds include CAS SciFinder Chemical Compound Database [27], Reaxys [28], and Pistachio [29]. The use of multiple data organization systems with differing naming conventions reduces the interoperability of molecular data and the ability to find, compare, and contrast molecules in a way that ensures that we are all communicating about the same molecule.

4. Semantic Materials Science for Molecular Data

We build a semantic data management (SDM) workflow to leverage the values of large scale molecular datasets in representation learning for application in EM, a small data domain. SDM relies on ontologies, knowledge graphs, and study protocols, following FAIR (findable, accessible, interoperable, reusable) principles, to reduce variability, ensure reproducibility, and allow data reuse and interoperability. An important aspect of the FAIRification of data is retaining the metadata about the sample, instrument specifications and settings, characterization specifications and settings, along with the collected data.

4.1. MDS-Onto: Molecule Identifiers

The first step in this data management is to create an ontology that aligns with the data and describes the critical metadata and provenance. We embedded the structural identifiers into a hierarchical method within MDS-Onto to allow for a searchable and structured database. The Materials Data

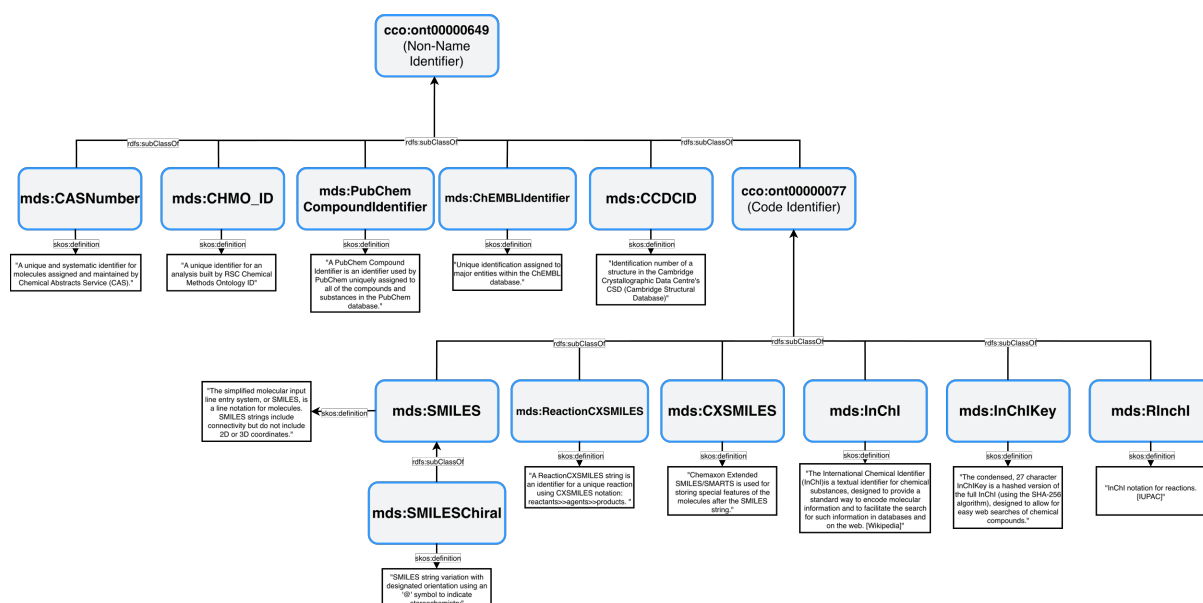


Figure 1: MDS-Onto terms for molecular identifiers and the hierarchy built.

Science Ontology (MDS-Onto) is an established ontology interoperable with the Basic Formal Ontology (BFO) that covers subfields in both materials science and chemistry [30]. MDS-Onto is a domain-level ontology for modeling studies and data in Materials Science and related disciplines [30]. It is built with connections to Basic Formal Ontology (BFO) & the Common Core Ontologies (CCO) [31, 32]. We expanded on MDS-Onto to build semantics of molecular data and helper terminology were pulled from BFO-compliant ontologies, Chemical Entities of Biological Interest (ChEBI) and Common Core Ontology (CCO). We defined and built a hierarchy of molecular identifiers shown in **Figure 1**: `mds:CASNumber`, `mds:CHMO_ID`, `mds:PubChemCompoundIdentifier`, `mds:ChEMBLIdentifier` are subclasses of `cco:00000649` (non-name identifier) and `mds:SMILES`, `mds:ReactionCXSMILES`, `mds:CXSMILES`, `mds:InChI`, `mds:InChIKey`, `mds:RInChI`, `mds:IUPACName`, and `mds:CompoundName`, are subclasses of `cco:ont0000077` (code identifier). The classes `mds:PubChemCompoundIdentifier` (PubChem), `mds:ChEMBLIdentifier` (ChEMBL), and `mds:CCDCID` (Cambridge Structural Database) were built into MDS-Onto to ensure molecules and their data from these databases are seamlessly integrated using MDS-Onto as the common ground while retaining the semantic relations with other identifiers such as SMILES and InChI. The rich identifiers for compounds ensure clarity and consistency (reduce ambiguity) when discussing molecules.

4.2. MDS-Onto: Functional Groups and Bond Types

While standard compound identifiers, such as SMILES or InChI, implicitly contain molecular topology, they often lack the explicit granular detail required for sophisticated machine learning applications. Incorporating specific ontological terms enables the explicit assignment of molecular domains, functional groups, and bonding characteristics. We expanded MDS-Onto to describe atom/bond counts and functional groups commonly encountered in the organic chemistry. We reused `obo:CHEBI_24432`, a chemical entity in Open Biological and Biomedical Ontology (OBO Foundry) compliant with BFO and sub-classed `mds:AtomCount` and `mds:FunctionalGroup` from chemical entity [33]. Counts of various types are defined further:

- `mds:CarbonBondedToCarbonAndTwoNitrogensCount`,
- `mds:OxygenBondedToCarbonAndNitrogenCount`,
- `mds:CarbonBondedToTwoCarbonsAndOxygenCount`.

These bond counts are commonly used as chemical descriptors in machine learning pipelines to represent

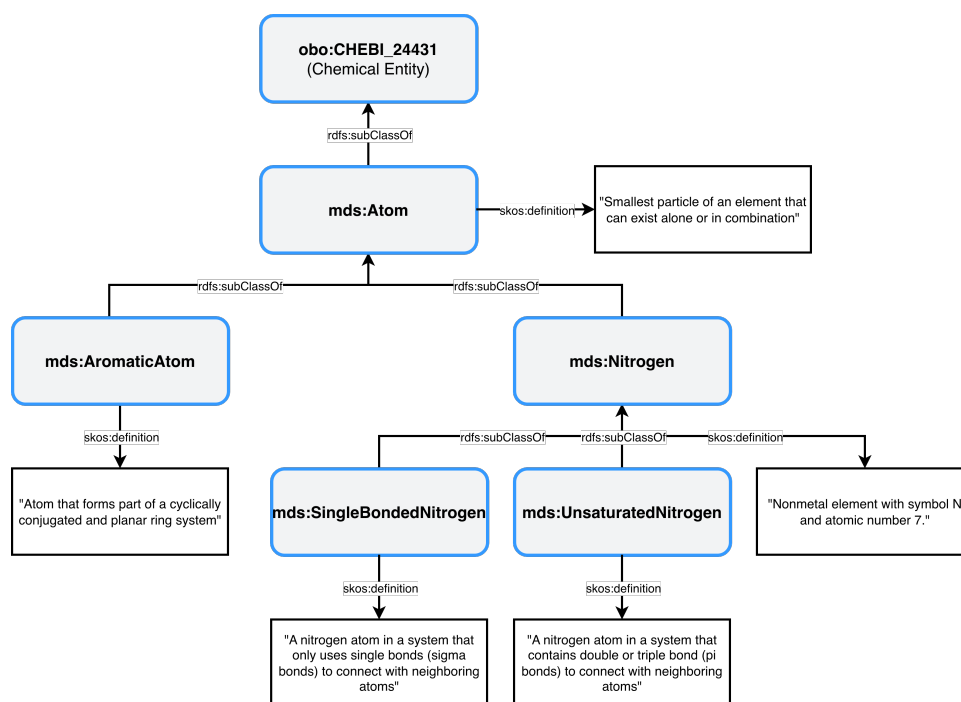


Figure 2: Ontology for atom, nitrogen atoms (single-bonded and unsaturated nitrogen), and aromatic atoms.

molecules with their structural information to correlate with their properties. Thus, they are critical to define and categorize in MDS-Onto to semantically describe molecules. An example of count terms are shown in **Figure 2**. To describe functional groups, there are 5 generic terms that encompass the most common functional groups in organic chemistry:

- `mds:CarbonbasedFunctionalGroup`,
- `mds:NitrogenbasedFunctionalGroup`,
- `mds:OxygenbasedFunctionalGroup`,
- `mds:HalogenbasedFunctionalGroup`,
- `mds:SulfurbasedFunctionalGroup`.

We defined the common functional groups as a subclass of one or more of these terms such as `mds:Ketone`, which is a subclass of `mds:CarbonbasedFunctionalGroup` and `mds:OxygenbasedFunctionalGroup` and `mds:Nitramine`, which is a subclass of `mds:CarbonbasedFunctionalGroup` and `mds:NitrogenbasedFunctionalGroup`. Examples of Nitrogen-based, Oxygen-based and Carbon-based functional groups are shown in **Figure 3** and **Figure 4**. Rather than solely relying on implicit information, we semantically pre-defined explicit feature mappings across different datasets of the features mentioned above. We apply this semantic grounding method across many historical datasets to unify compound identification, something we have considered a necessary challenge to overcome. The identification can move from typical molecular names that can create confusion with inconsistencies and focus on explicit counts.

4.3. Data Curation: RDF Linked Data for Knowledge Management

We built a large dataset of over 3.7 million existing molecules from open-source databases: ChEMBL, PubChem, and Open Reaction Database (ORD). Naming systems and the schemas of these databases (e.g. Compound Identifier or CID from PubChem, CSD refcodes, and ChEMBLID) are incorporated into MDS-Onto to ensure interoperability and track provenance of any molecule in our database as described previously. We retained SMILES strings, chemical/molecular formulas, InChI strings, compound names, and other identifying data. **Table 1** displays the density and sparsity of the identifying data within

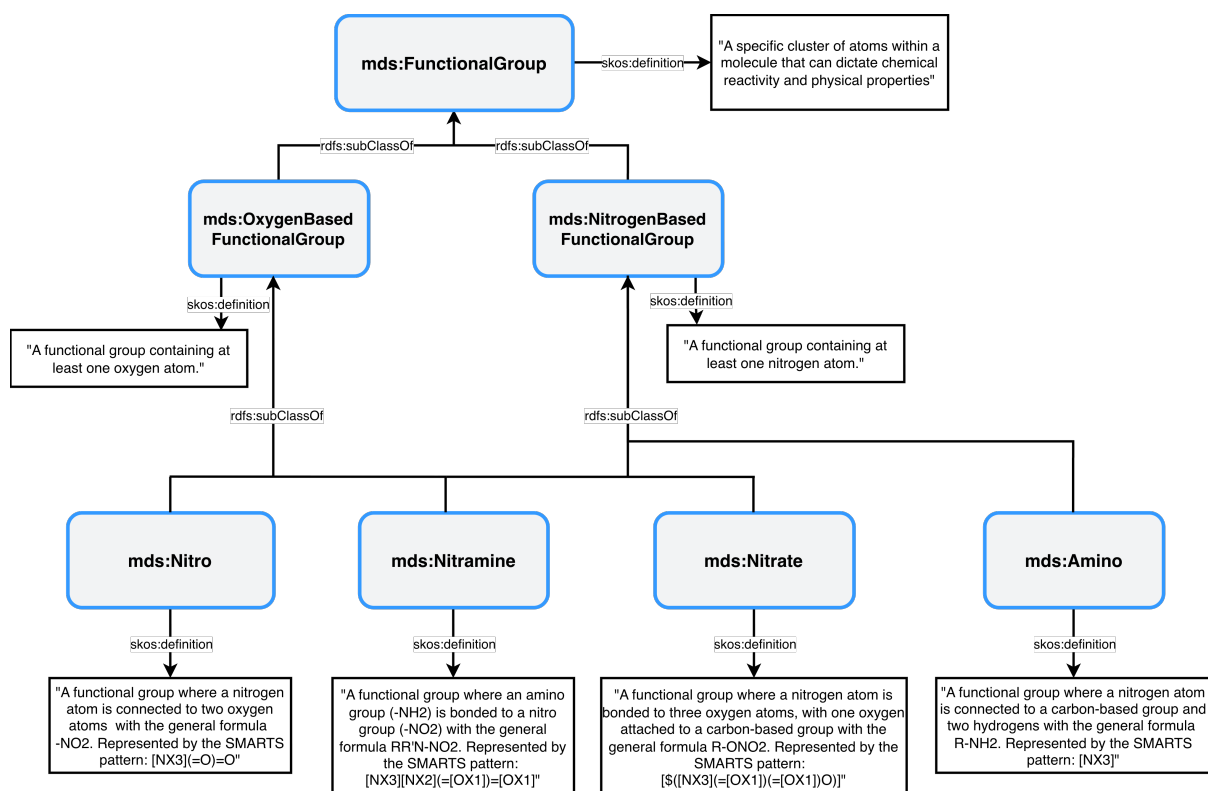


Figure 3: Examples of nitrogen-based functional groups: mds:Nitro ($-\text{NO}_2$), mds:Nitramine ($\text{RR}'\text{N}-\text{NO}_2$), mds:Nitrate ($\text{R}-\text{ONO}_2$), and mds:Amino ($\text{R}-\text{NH}_2$).

Table 1

Molecular identifiers and counts in the integrated dataset.

Columns	Rows With Data	Missing/Empty Rows	% Rows Filled
ml#	3,724,783	0	100%
SMILES	3,724,783	0	100%
CompoundName	701,572	3,023,211	18.84%
InChI	3,724,783	0	100%
InChIKey	3,724,777	6	100%
IUPACName	793,847	2,930,936	21.31%
PubChem Compound Identifier	792,986	2,931,797	21.29%
ChEMBL Identifier	2,700,015	1,024,768	72.49%
CanonicalSMILES	3,724,783	0	100%

our 3.7 million molecules. In the curated dataset, 59,000 molecules have both a PubChem compound identifier (CID) and a ChEMBL identifier and 291,000 molecules have neither, meaning that around 3.2 million molecules have at least one of the two identifiers. All of SMILES strings are validated and canonical SMILES are generated using RDKit to de-duplicate the combined dataset. If a molecule has multiple valid SMILES strings, all of them are stored to ensure future validation of new molecules are comprehensive. To retrieve data from each of the sources, we tailored to each of the Python APIs available for them. Each of these sources has its own identifying method for its compounds; this is why we opted to build a script to validate all molecules in our system to ensure consistency. We used `mds:SMILES`, `mds:InChI`, `mds:PubChemCompoundIdentifier`, `mds:ChEMBLIdentifier` to validate, de-duplicate, and add missing metadata to molecules. We transformed molecular data from various .csv files into RDF graphs of linked data using the FAIRLinked, a Python open-source package developed in-house to facilitate data transformation to JSON-LD [34]. This molecular data is seamlessly integrated with the Chemical Reaction Knowledge Graph (ChemRxnKG), which is hosted on Open Science Foundry

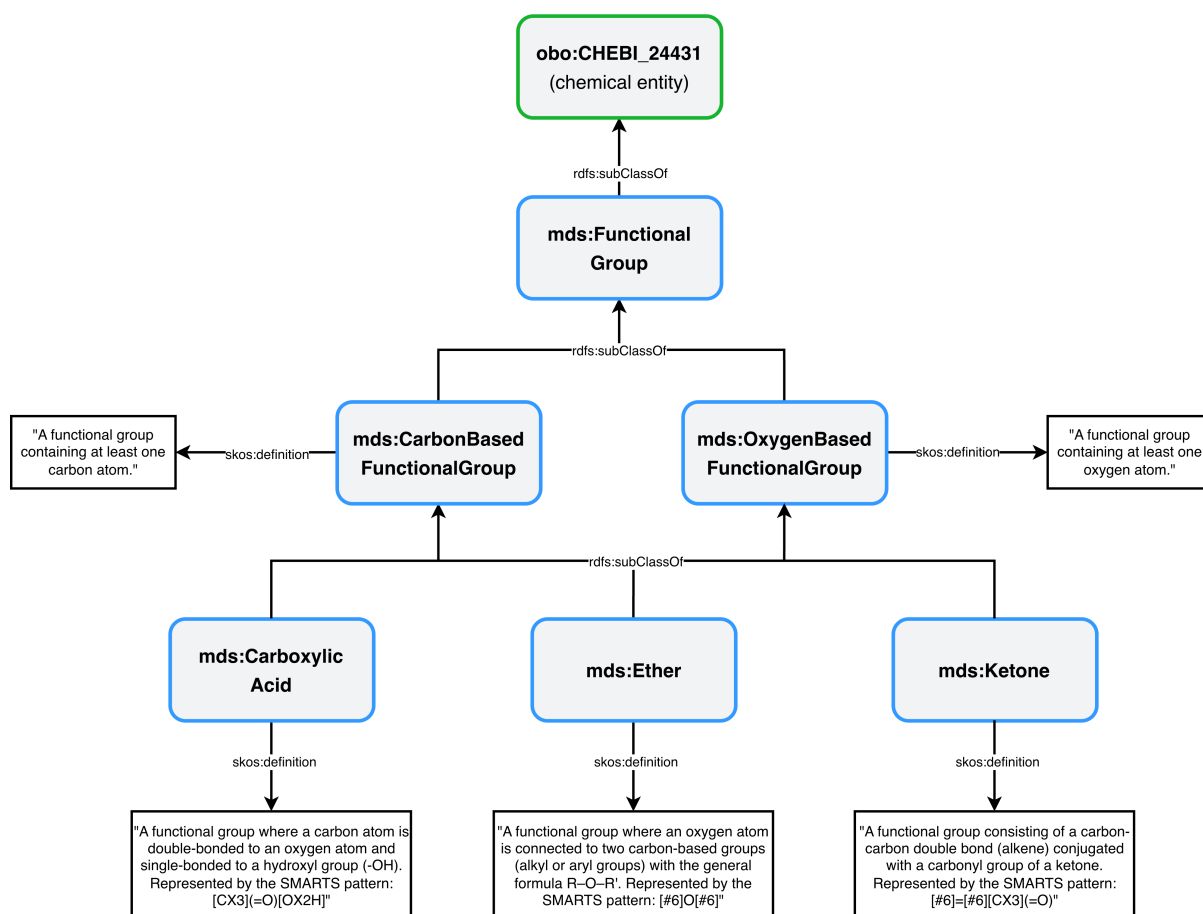


Figure 4: Examples of carbon-based and oxygen-based functional groups: `mds:CarboxylicAcid` ($-\text{C}(=\text{O})\text{O}$), `mds:Ether` ($-\text{COC}-$), and `mds:Ketone` ($-\text{C}(=\text{O})\text{C}-$).

(OSF), without breaking the existing schema since both are built with MDS-Onto [35]. MSD-Onto is expanded to structure the data as previously described and enable semantic reasoning. This allowed for easier classification of the compounds and their properties instead of having to treat each of the databases as isolated sources. The molecular data and identifiers are ingested into HBase, a NoSQL database for big data developed under the Hadoop ecosystem. The molecules are annotated with the functional groups and the types along with the counts of bonds present to enrich the semantics of the molecules. The curated molecular dataset is a validation and lookup table for incoming molecules, chemical reaction data, or any materials or sample studied in any materials or chemistry investigation.

This comprehensive dataset can be used in a wide range of molecular and materials studies. An example includes energetic materials (EM) property prediction and discovery. To explore potential energetic molecules in our dataset, we generated Morgan Fingerprints using RDKit, an open-source Python package for cheminformatics, for 3.7 million molecules in the database and use t-distributed Stochastic Neighbor Embedding (t-SNE) plots to visualize the distribution of the fingerprints color-coded by the types and counts of 5 nitrogen-based functional groups, shown in **Figure 5**: `mds:Nitrate` ($\text{CON}(=\text{O})=\text{O}$), `mds:Azide` ($\text{CN}=[\text{N}^+]=[\text{N}-]$), `mds:Nitro` ($\text{C}[\text{N}^+](=\text{O})[\text{O}-]$), `mds:Nitramine` ($\text{N}-\text{N}(=\text{O})=\text{O}$), and `mds:Azo` ($\text{C}-\text{N}=\text{N}-\text{C}$) [36]. Morgan Fingerprints are bit vectors indicating the presence (1) and absence (0) of unique atomic environments within a molecule, thus, are structural representations of molecules. The colored cluster in **Figure 5** show that molecules with `mds:Nitrate`, which is the most common of the 5 functional groups, and mixed/multi-type functional groups are clustered together. Molecules with other types are scattered throughout. This is expected as structurally similar molecules, indicated by the fingerprints, are closer together and structurally dissimilar molecules are further apart. The applicability of the curated dataset is not limited to 5 functional groups and EM research questions,

rather, it can be used to answer research questions concerning for example molecular scaffolds in biochemistry.

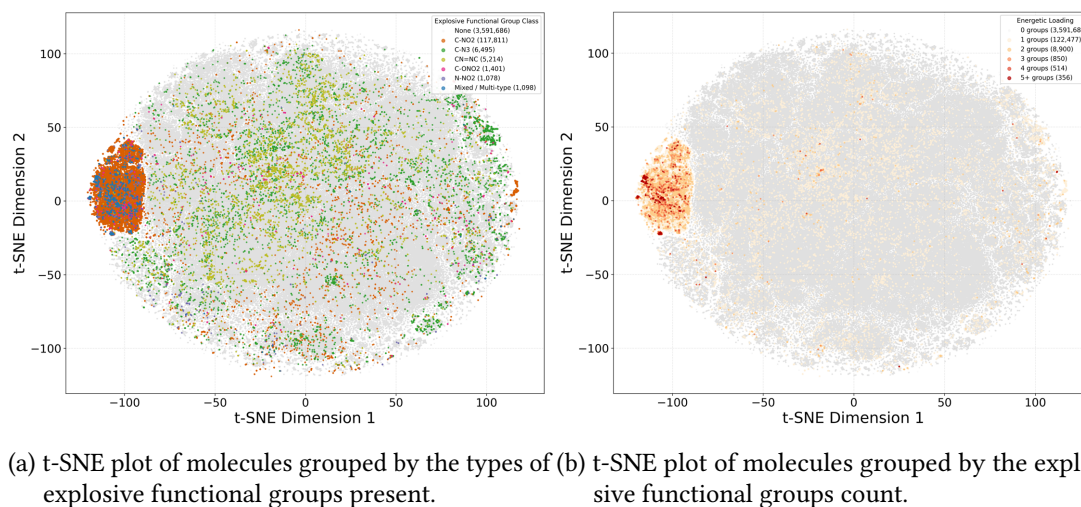


Figure 5: t-SNE plot of 3.7 million molecules from ChEMBL, ORD, and PubChem by nitrogen-based functional groups: mds:Nitrate ($\text{CON}(=\text{O})=\text{O}$), mds:Azide ($\text{CN}=[\text{N}^+]=[\text{N}^-]$), mds:Nitro ($\text{C}[\text{N}^+](=\text{O})[\text{O}^-]$), mds:Nitramine ($\text{N}-\text{N}(=\text{O})=\text{O}$), and mds:Azo ($\text{C}-\text{N}=\text{N}-\text{C}$).

5. Case Study: Large Scale Learning of Molecular Data

We demonstrate a simple example of large scale graph-based learning on the integrated molecular dataset as a curated molecular dataset of existing molecules from multiple databases with proper annotation of functional groups and bond counts based on an ontology schema does not currently exist to our knowledge. A series of 6 different graph-based encoders from simple to advanced were used with standard PyG out-of-the-box scripts to generate molecular embeddings: graph convolutional network (GCN), graph attention network (GAT), graph isomorphism network (GIN), directed message-passing neural network (DMPNN), graph transformer, and hybrid DMPNN-Transformer [37]. The encoders take featurized molecular graphs and the connectivity via an adjacency matrix as inputs. The atomic features include atom type, degree, chirality, valence electrons, hybridization, etc. and the bond features include single, double, triple, or aromatic. The molecules are featurized with functional group tags to enrich the graphs. Using Principal Component Analysis (PCA) and t-SNE, we can visualize the distribution the molecular embeddings from each of the models color-coded by the type of functional groups present as shown in **Figure 6**.

The molecules with the nitrogen-based functional groups of interest are scattered throughout. Molecular embeddings, unlike fingerprints that only contain 0s and 1s, contain more information about the atoms and bonds in molecules as features, not just structural information (i.e. atomic environments). Qualitative analysis of the t-SNE plots reveals that the distribution of the molecular embeddings based on the mds:Nitrate and other functional groups are not clustered similarly to the fingerprints regardless of the type of encoder. Structurally similar molecules are not close together indicating that they have dissimilar atomic and bond features despite structural similarities. Quantitative analysis at scale of molecular embeddings, functional groups, Morgan Fingerprints, and different encoders is required to further uncover the underlying patterns. An integrated and curated large scale molecular database made possible with MDS-Onto is necessary to investigate molecular representations, deep learning, and benchmarking for prediction and discovery tasks.

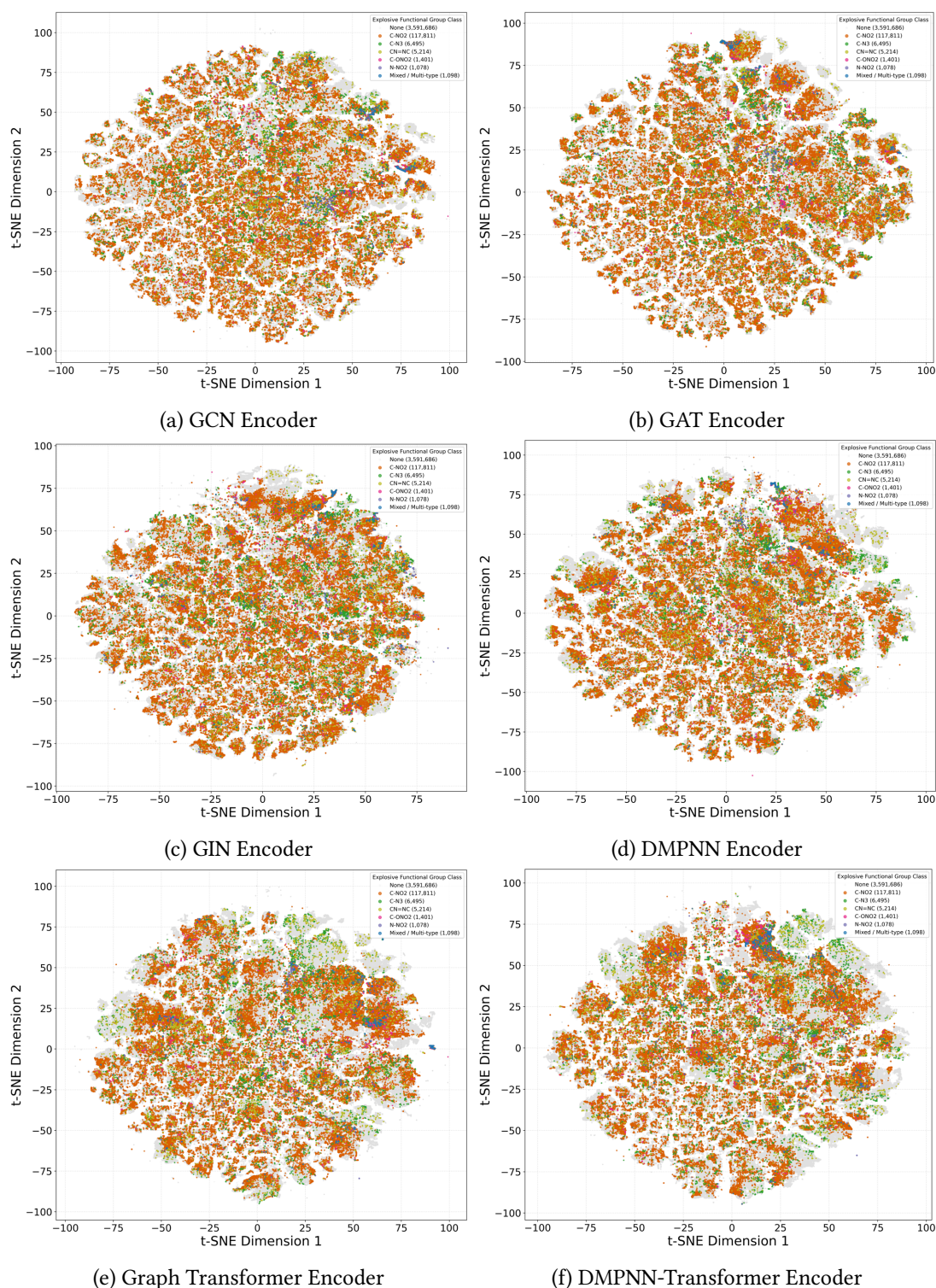


Figure 6: t-SNE plots of molecular embeddings of 3.7 million molecules generated with different encoders by nitrogen-based functional groups: mds:Nitrate ($\text{CON}(\text{=O})\text{O}$), mds:Azide ($\text{CN}=[\text{N}^+]=[\text{N}-]$), mds:Nitro ($\text{C}[\text{N}^+](\text{=O})[\text{O}-]$), mds:Nitramine ($\text{N}-\text{N}(\text{=O})\text{O}$), and mds:Azo ($\text{C}-\text{N}=\text{N}-\text{C}$).

6. Conclusion

Molecular and materials discovery using AI/ML is limited by many data challenges and biases. A comprehensive effort to unify molecules from open-source experimental molecular databases and resolve discrepancies using ontology for large scale molecular and materials studies does not exist.

In this work, we introduce a workflow for implementing semantic data management in chemical data for a more standardized organization method to FAIRify and store data and metadata. We have established a continuously growing and open ontology to map molecular and energetic data creating a framework for combining historical and new chemical datasets implementing semantic materials science to organize and uniquely identify existing molecules. This makes way for the use of larger datasets within machine learning for a more comprehensive understanding of molecular relationships and representation. The workflow presented ensures reproducibility and interoperable data for downstream ML tasks in molecular discovery and property prediction. We built an semantic-rich database of molecules sourced from multiple open-source molecular datasets and will continue to expand this database. The database enables researchers to perform semantic queries of molecules in-house without relying on navigating multiple outdated APIs and answer. This database will be shared freely with any researcher interested, inciting collaboration and expediting scientific discovery. Exploration of various molecular representations in deep learning with ontologies is an area of active research to develop better understanding of the learning process in DL models.

Acknowledgments

This material is based upon research in the Materials Data Science for Stockpile Stewardship Center of Excellence (MDS3-COE), and supported by the U.S. Department of Energy's National Nuclear Security Administration under Award Number(s) DE-NA0004104. This work is supported by AI4NS project with the Pantex Plant under the contract number PTX01-0000110811. This work made use of the High Performance Computing Resource in the Core Facility for Advanced Research Computing at Case Western Reserve University.

Data Statement

Data in the ChemRxnKG as RDF triples is available on Open Science Foundation with a persistent DOI: <https://osf.io/5ecsf>. MDS-Onto Open is an open-source project: <https://cwrusdle.bitbucket.io/>. The molecular dataset as graph data is available via request.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

- [1] N. Pallikkavaliyaveetil, S. Chandrasekaran, Small data, big challenges: Machine- and deep-learning strategies for data-limited drug discovery, *Advanced drug delivery reviews* 229 (2026) 115762. doi:10.1016/j.addr.2025.115762.
- [2] S. Kim, J. Chen, T. Cheng, A. Gindulyte, J. He, S. He, Q. Li, B. A. Shoemaker, P. A. Thiessen, B. Yu, L. Zaslavsky, J. Zhang, E. E. Bolton, PubChem 2025 update, *Nucleic Acids Research* 53 (2025) D1516–D1525. doi:10.1093/nar/gkae1059.
- [3] B. Zdrazil, E. Felix, F. Hunter, E. J. Manners, J. Blackshaw, S. Corbett, M. de Veij, H. Ioannidis, D. M. Lopez, J. F. Mosquera, M. P. Magarinos, N. Bosc, R. Arcila, T. Kizilören, A. Gaulton, A. P. Bento, M. F. Adasme, P. Monecke, G. A. Landrum, A. R. Leach, The ChEMBL Database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods, *Nucleic Acids Research* 52 (2024) D1180–D1192. doi:10.1093/nar/gkad1004.

- [4] C. R. Groom, I. J. Bruno, M. P. Lightfoot, S. C. Ward, The Cambridge Structural Database, *Acta Crystallographica Section B: Structural Science, Crystal Engineering and Materials* 72 (2016) 171–179. doi:10.1107/S2052520616003954.
- [5] S. M. Kearnes, M. R. Maser, M. Wleklinski, A. Kast, A. G. Doyle, S. D. Dreher, J. M. Hawkins, K. F. Jensen, C. W. Coley, The open reaction database, *Journal of the American Chemical Society* 143 (2021) 18820–18826. URL: <https://doi.org/10.1021/jacs.1c09820>. doi:10.1021/jacs.1c09820. arXiv:<https://doi.org/10.1021/jacs.1c09820>, PMID: 34727496.
- [6] J. J. Irwin, K. G. Tang, J. Young, C. Dandarchuluun, B. R. Wong, M. Khurelbaatar, Y. S. Moroz, J. Mayfield, R. A. Sayle, Zinc20—a free ultralarge-scale chemical database for ligand discovery, *Journal of Chemical Information and Modeling* 60 (2020) 6065–6073. doi:10.1021/acs.jcim.0c00675.
- [7] T. Fink, H. Bruggesser, J.-L. Reymond, Virtual exploration of the small-molecule chemical universe below 160 daltons, *Angewandte Chemie International Edition* 44 (2005) 1504–1508. doi:10.1002/anie.200462457.
- [8] Y. Buehler, J.-L. Reymond, Molecular framework analysis of the generated database gdb-13s, *Journal of Chemical Information and Modeling* 63 (2022) 484–492. doi:10.1021/acs.jcim.2c01107.
- [9] Y. Buehler, J.-L. Reymond, Expanding bioactive fragment space with the generated database gdb-13s, *Journal of Chemical Information and Modeling* 63 (2023) 6239–6248. doi:10.1021/acs.jcim.3c01096.
- [10] Y. Buehler, J.-L. Reymond, A view on molecular complexity from the gdb chemical space, *Journal of Chemical Information and Modeling* 65 (2025) 8405–8410. doi:10.1021/acs.jcim.5c00334.
- [11] W. L. Hamilton, R. Ying, J. Leskovec, Representation Learning on Graphs: Methods and Applications, arXiv:1709.05584 [cs] (2018). URL: <http://arxiv.org/abs/1709.05584>.
- [12] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. G. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. C. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, B. Mons, The FAIR Guiding Principles for scientific data management and stewardship, *Scientific Data* 3 (2016) 160018. doi:10.1038/sdata.2016.18.
- [13] A. Jacobsen, R. de Miranda Azevedo, N. Juty, D. Batista, S. Coles, R. Cornet, M. Courtot, M. Crosas, M. Dumontier, C. T. Evelo, C. Goble, G. Guizzardi, K. K. Hansen, A. Hasnain, K. Hettne, J. Heringa, R. W. Hooft, M. Imming, K. G. Jeffery, R. Kaliyaperumal, M. G. Kersloot, C. R. Kirkpatrick, T. Kuhn, I. Labastida, B. Magagna, P. McQuilton, N. Meyers, A. Montesanti, M. van Reisen, P. Rocca-Serra, R. Pergl, S.-A. Sansone, L. O. B. da Silva Santos, J. Schneider, G. Strawn, M. Thompson, A. Waagmeester, T. Weigel, M. D. Wilkinson, E. L. Willighagen, P. Wittenburg, M. Roos, B. Mons, E. Schultes, FAIR Principles: Interpretations and Implementation Considerations, *Data Intelligence* 2 (2020) 10–29. doi:10.1162/dint_r_00024.
- [14] Q. D. Tran, E. I. Barcelos, L. S. Bruckman, Designing Data-Centric Study Protocols Guided by FAIR Principles, *Interdisciplinary Information Sciences* 31 (2025) 75–82. doi:10.4036/iis.2025.A.07.
- [15] S. Skonieczny, The IUPAC Rules for Naming Organic Molecules, *Journal of Chemical Education* 83 (2006) 1633. doi:10.1021/ed083p1633.
- [16] D. Weininger, SMILES, a chemical language and information system. Introduction to methodology and encoding rules, *Journal of Chemical Information and Computer Sciences* 28 (1988) 31–36. doi:10.1021/ci00057a005.
- [17] ChemAxon, Chemaxon SMILES extensions, https://docs.chemaxon.com/latest/formats_chemaxon-smiles-extensions.html, 2024. Accessed: 2024-03-01.
- [18] S. R. Heller, A. McNaught, I. Pletnev, S. Stein, D. Tchekhovskoi, InChI, the IUPAC International Chemical Identifier, *Journal of Cheminformatics* 7 (2015) 23. doi:10.1186/s13321-015-0068-4.
- [19] M. Krenn, Q. Ai, S. Barthel, N. Carson, A. Frei, N. C. Frey, P. Friederich, T. Gaudin, A. A. Gayle, K. M.

- Jablonka, R. F. Lameiro, D. Lemm, A. Lo, S. M. Moosavi, J. M. Nápoles-Duarte, A. Nigam, R. Pollice, K. Rajan, U. Schatzschneider, P. Schwaller, M. Skreta, B. Smit, F. Strieth-Kalthoff, C. Sun, G. Tom, G. F. von Rudorff, A. Wang, A. D. White, A. Young, R. Yu, A. Aspuru-Guzik, Selfies and the future of molecular string representations, *Patterns* 3 (2022) 100588. URL: <https://www.sciencedirect.com/science/article/pii/S2666389922002069>. doi:<https://doi.org/10.1016/j.patter.2022.100588>.
- [20] T. Gokey, D. L. Mobley, Hierarchical clustering of chemical space using binary-encoded smarts for building data-driven chemical perception models, *ChemRxiv* 2023 (2023). URL: <https://chemrxiv.org/doi/abs/10.26434/chemrxiv-2023-v969f-v3>. doi:10.26434/chemrxiv-2023-v969f-v3. arXiv:<https://chemrxiv.org/doi/pdf/10.26434/chemrxiv-2023-v969f-v3>.
- [21] S. Aykent, T. Xia, Gotennet: Rethinking efficient 3d equivariant graph neural networks, 2024. URL: <https://openreview.net/forum?id=5wxCQDtbMo>.
- [22] K. Wu, B. Natarajan, L. Morkowchuk, M. Krein, C. M. Breneman, Chapter 16 - From Drug Discovery QSAR to Predictive Materials QSPR: The Evolution of Descriptors, Methods, and Models, Butterworth-Heinemann, Oxford, 2013, p. 385–422. URL: <http://www.sciencedirect.com/science/article/pii/B9780123943996000163>, 00000.
- [23] M. Krenn, F. Häse, A. Nigam, P. Friederich, A. Aspuru-Guzik, Self-Referencing Embedded Strings (SELFIES): A 100% robust molecular string representation, *Machine Learning: Science and Technology* 1 (2020) 045024. doi:10.1088/2632-2153/aba947, arXiv:1905.13741 [cs.LG].
- [24] M. Davies, M. Nowotka, G. Papadatos, N. Dedman, A. Gaulton, F. Atkinson, L. Bellis, J. P. Overington, ChEMBL web services: streamlining access to drug discovery data and utilities, *Nucleic Acids Research* 43 (2015) W612–W620. doi:10.1093/nar/gkv352.
- [25] S. Jupp, J. Malone, J. Bolleman, M. Brandizi, M. Davies, L. Garcia, A. Gaulton, S. Gehant, C. Laibe, N. Redaschi, S. M. Wimalaratne, M. Martin, N. Le Novère, H. Parkinson, E. Birney, A. M. Jenkinson, The EBI RDF platform: linked open data for the life sciences, *Bioinformatics* 30 (2014) 1338–1339. doi:10.1093/bioinformatics/btt765.
- [26] P. de Matos, A. Dekker, M. Ennis, J. Hastings, K. Haug, S. Turner, C. Steinbeck, ChEBI: a chemistry ontology and database, *Journal of Cheminformatics* 2 (2010) P6. doi:10.1186/1758-2946-2-S1-P6.
- [27] S. W. Gabrielson, Scifinder, *Journal of the Medical Library Association: JMLA* 106 (2018) 588–590. doi:10.5195/jmla.2018.515.
- [28] Elsevier, Reaxys® — chemistry data and ai to optimize small molecule discovery, 2026. URL: <https://www.elsevier.com/products/reaxys>.
- [29] R. A. Sayle, D. Sayle, J. Mayfield, I. Lagerstedt, R. Pirie, L. Heinzke, Pistachio: Search and faceting of large reaction databases – nextmove software, 2017. URL: <https://nextmovesoftware.com/blog/2017/12/11/pistachio-search-and-faceting-of-large-reaction-databases/>.
- [30] B. P. Rajamohan, A. C. H. Bradley, V. D. Tran, J. E. Gordon, H. W. Caldwell, R. Mehdi, G. Ponon, Q. D. Tran, O. Dernek, J. Kaltenbaugh, B. G. Pierce, R. Wieser, W. Yue, K. Lin, J. Kambo, C. Lopez, A. Nihar, D. J. Savage, D. W. Brown, H. Sharma, B. Giera, P. K. Tripathi, Y. Wu, M. Li, K. O. Davis, L. S. Bruckman, E. I. Barcelos, R. H. French, Materials Data Science Ontology(MDS-Onto): Unifying Domain Knowledge in Materials and Applied Data Science, *Scientific Data* 12 (2025) 628. doi:10.1038/s41597-025-04938-5.
- [31] R. Arp, B. Smith, A. D. Spear, Building Ontologies with Basic Formal Ontology, MIT Press, 2015. URL: <https://direct.mit.edu/books/monograph/4044/Building-Ontologies-with-Basic-Formal-Ontology>.
- [32] M. Jensen, G. D. Colle, S. Kindya, C. More, A. P. Cox, J. Beverley, The Common Core Ontologies (2024). URL: <http://arxiv.org/abs/2404.17758>. doi:10.48550/arXiv.2404.17758, arXiv:2404.17758 [cs].
- [33] B. Smith, M. Ashburner, C. Rosse, J. Bard, W. Bug, W. Ceusters, L. J. Goldberg, K. Eilbeck, A. Ireland, C. J. Mungall, N. Leontis, P. Rocca-Serra, A. Ruttenberg, S.-A. Sansone, R. H. Scheuermann, N. Shah, P. L. Whetzel, S. Lewis, The obo foundry: coordinated evolution of ontologies to support biomedical data integration, *Nature Biotechnology* 25 (2007) 1251–1255. doi:10.1038/nbt1346.
- [34] V. D. Tran, B. Lee, R. Lamba, H. Dirks, Q. D. Tran, B. P. Rajamohan, O. Dernek, L. S. Bruckman, Y. Wu, E. I. Barcelos, R. H. French, FAIRlinked: Data FAIRification Tools for Materials Data Science,

Journal of Open Source Software 11 (2026) 9628. doi:10.21105/joss.09628.

- [35] Q. Tran, H. Schreiber, B. Lee, O. Schessler, L. S. Bruckman, R. French, Chemrxnkg: Semantic data management for historical and sparse chemical reaction data (2026). URL: <https://osf.io/5ecsf>.
- [36] G. Landrum, V. Scalfani, C. Kuenneth, Rdkit: Open-source cheminformatics and machine learning, 2026. URL: <https://www.rdkit.org/docs/api-docs.html>.
- [37] M. Fey, J. Sunil, A. Nitta, R. Puri, M. Shah, B. Stojanovič, R. Bendias, A. Barghi, V. Kocijan, Z. Zhang, X. He, J. E. Lenssen, J. Leskovec, Pyg 2.0: Scalable learning on real world graphs (2025). URL: <http://arxiv.org/abs/2507.16991>. doi:10.48550/arXiv.2507.16991, arXiv:2507.16991 [cs].