Data Sources Document

Superconductor Formula VAE Project

Last Updated: January 19, 2026

1. Purpose

This document outlines all datasets to be integrated into the VAE codebase for superconductor formula generation. The capstone requires gigabyte-scale data, so we combine multiple sources. Each dataset serves a specific purpose in training.

2. Glossary of Abbreviations

This section defines abbreviations used throughout the document for team members new to superconductor research.

Superconductor Terms

- **Tc (Critical Temperature):** The temperature below which a material becomes superconducting (zero electrical resistance). Measured in Kelvin (K). Higher Tc = more useful.
- **GPa (Gigapascals):** Unit of pressure. Some materials only superconduct under extreme pressure (e.g., hydrogen-rich compounds at 100-500 GPa).
- **High-Tc:** Superconductors with critical temperatures above ~30K, typically cuprates (copper-oxide compounds) or iron-based materials.

Magnetic Classification Terms

- **FM (Ferromagnetic):** Materials where atomic magnetic moments align parallel (e.g., iron, cobalt). Generally incompatible with superconductivity.
- **AFM (Antiferromagnetic):** Materials where adjacent magnetic moments align anti-parallel, canceling out. Some antiferromagnets neighbor superconducting phases.
- **NM (Non-magnetic):** Materials with no ordered magnetic structure.
- **Curie Temperature:** Temperature above which a ferromagnet loses its permanent magnetism.

- **Néel Temperature:** Temperature above which an antiferromagnet loses its ordered magnetic structure.

Computational Chemistry Terms

- **DFT (Density Functional Theory):** A quantum mechanical method for calculating material properties from first principles. Most large materials databases use DFT-computed values.
- **Formation Energy:** Energy released/required to form a compound from its elements. Negative = stable compound.
- **Band Gap:** Energy difference between valence and conduction bands. Zero band gap = metallic (prerequisite for superconductivity).
- **Phonons:** Quantized lattice vibrations. Electron-phonon coupling is the mechanism behind conventional superconductivity.

File Format Terms

- **CIF (Crystallographic Information File):** Standard file format for crystal structures containing lattice parameters and atomic positions.
- **POSCAR:** Crystal structure file format used by VASP (a popular DFT software). Contains lattice vectors and atomic coordinates.
- **RDF (Resource Description Framework):** A data format for representing linked data. NIMS provides SuperCon data in RDF format.

Database Abbreviations

- **NIMS:** National Institute for Materials Science (Japan). Maintains the SuperCon database.
- **MDR:** Materials Data Repository (NIMS's data platform).
- **OQMD:** Open Quantum Materials Database. ~1.3 million DFT-calculated materials.
- **AFLOW:** Automatic FLOW for Materials Discovery. Large-scale computational materials database.
- **NEMAD:** Novel Electronic and Magnetic Materials Database. Contains magnetic property data.
- **UCI:** University of California, Irvine (hosts the ML repository with pre-processed SuperCon data).

Machine Learning Terms

- **VAE (Variational Autoencoder):** A generative neural network that learns a compressed latent representation of data and can generate new samples.
- **GAN (Generative Adversarial Network):** A generative model using two competing networks (generator vs. discriminator).
- **DDPM (Denoising Diffusion Probabilistic Model):** A generative model that learns to reverse a gradual noising process.
- **CDVAE:** Crystal Diffusion Variational Autoencoder. Combines VAE with diffusion for 3D crystal generation.
- **Magpie:** Materials-Agnostic Platform for Informatics and Exploration. Generates 145 compositional features from chemical formulas.
- **Transfer Learning:** Pre-training a model on a large dataset, then fine-tuning on a smaller target dataset.
- **Contrastive Learning:** Training a model to distinguish between similar and dissimilar examples (e.g., superconductors vs. magnetic materials).

Other Abbreviations

- **CC BY 4.0:** Creative Commons Attribution 4.0 license. Free to use with attribution.
- **SQL:** Structured Query Language. OQMD is distributed as an SQL database dump.
- **MVP:** Minimum Viable Product. The initial working version of the system.

3. Priority Summary

Phase 1 (Required for MVP): MDR SuperCon + OQMD

Phase 2 (Recommended): SuperCon2 + NEMAD

Phase 3 (Future Extension): JARVIS-DFT + Superhydra

4. Primary Superconductor Databases

These are the core datasets containing known superconductors with critical temperature (Tc) values.

Table 1: Primary Superconductor Databases

Integration Notes for SuperCon:

- MDR SuperCon is the canonical source used by most published methods (ScGAN, SuperDiff, Stanev et al.)
- License: CC BY 4.0 (free to use with attribution)
- SuperCon2 adds pressure data which is critical for high-pressure hydride superconductors
- **Important:** Do NOT label non-superconductors as Tc=0. Use NA/null for unknown Tc values.

5. Scale-Up Datasets (GB Requirement)

The capstone requires gigabyte-scale data. SuperCon alone is ~20MB. These general materials databases provide the necessary scale.

Table 2: Scale-Up Databases for GB Requirement

Recommended Approach:

Use OQMD as the primary scale-up dataset. This is the cleanest path to GB-scale data and mirrors what ScGAN did successfully.

- Download: oqmd.org/download (single ~4GB file)
- Format: SQL dump or REST API
- Training strategy: Pre-train VAE on OQMD for general materials representation, then fine-tune on SuperCon for Tc prediction

6. Contrastive Learning Dataset

Training on both superconductors AND magnetic materials helps the VAE learn the full magnetic-superconducting phase space.

Table 3: Contrastive Learning Dataset (Magnetic Materials)

Why NEMAD Matters:

- Superconductors and magnetic materials represent opposing phase behaviors
- The VAE can learn what makes a material superconducting vs. ferromagnetic/antiferromagnetic
- Access: www.nemad.org
- Reference: Itani et al. 2024/2025

7. Future Extension Datasets (Phase 3)

These datasets add 3D crystal structures and high-pressure data for more advanced modeling.

Table 4: Future Extension Datasets (3D Structures & High Pressure)

8. What Prior Methods Used

For reference, here is what published generative models used:

Table 5: Datasets Used by Prior Generative Methods

Key observation: All composition-level methods (ScGAN, SuperDiff) use vector encodings. Our sequence-based approach with pointer-generator is novel in this space.

9. Integration Checklist

- **MDR SuperCon:** Download CSV, parse formulas, extract Tc values. Handle missing Tc as null.
- **OQMD:** Download SQL dump (~4GB). Extract formula + formation energy. Create train/val/test splits.
- **NEMAD:** Download from nemad.org. Extract formula + magnetic classification. Merge with superconductor data.
- **Unify schema:** Create common format with fields: formula, Tc (nullable), is_superconductor (bool), magnetic_class (nullable), source_db.
- **Compute Magpie features:** Use matminer/Magpie to generate 145 compositional descriptors for each formula.
- **Holdout set:** Reserve specific high-Tc compounds (e.g., $HgBa_2Ca_2Cu_3O_8$) for generalization testing.

10. Questions?

Reach out if you need clarification on any dataset, access issues, or integration priorities.

Database	Records	Key Properties	Access URL
MDR SuperCon	~32,000		mdr.nims.go.jp/collections/5712mb227

		Tc, formula, material class (oxide/metallic/organic)	
SuperCon2	40,324	Tc, pressure, measurement method, doping, substrate	mdr.nims.go.jp (search SuperCon2)
UCI Superconductivity	21,263	81 pre-computed features + Tc	archive.ics.uci.edu/dataset/464
Kaggle Cleaned SuperCon	30,000+	Processed from NIMS RDF	kaggle.com/datasets/sautkin/cleaned-supercon-from-nims-rdf-1-2

Database	**Records / Disk**	**Contents**	**Literature Usage**
OQMD	1.3M / ~4 GB	DFT formation energies, lattice params, atomic positions	ScGAN used for pre-training before transfer to SuperCon
Materials Project	150k / 1-2 GB	Structures, band gaps, formation energies	CDVAE training, property prediction benchmarks
AFLOW	3.5M / 40+ TB	Band structures, thermochemical properties	Query specific slices (not full download)

Database	**Records**	**Properties**
NEMAD	67,573	FM/AFM/NM classification, Curie temperature, Néel temperature, magnetization, coercivity, susceptibility

Database	**Records**	**Contents**	**Use Case**
JARVIS-DFT	~40,000	3D structures (POSCAR), band structure, phonons	Extend VAE to structure-aware generation
Superhydra	~880		Test set for H_3S-type discovery

		High-pressure hydrides (0-500 GPa) with CIF structures	
NIST High-P Hydrides	~900	DFT at 0, 100, 200, 300, 500 GPa	Pressure-dependent Tc modeling

Method	**Year**	**Dataset(s)**	**Approach**
Stanev et al.	2018	SuperCon (~21k)	Magpie features + Random Forest (foundational)
ScGAN	2023	OQMD → SuperCon	GAN with transfer learning, vector encoding
SuperDiff	2024	SuperCon (~21k)	DDPM with ILVR conditioning, vector encoding
Guided Diffusion (Prakash)	2025	7,183 w/ structures	DiffCSP fine-tuning, 3D crystal generation
CDVAE	2022	MP-20 (45k), Perov-5 (19k)	Crystal diffusion VAE, 3D structures