Superconductor Formula VAE: A Fraction-Aware Variational Autoencoder with Pointer–Generator Decoding for Generative Superconductor Discovery

Giorgia Foresta, Ninelia Gharakhani, James Conde, Albab Newaz

Superconducting materials discovery remains challenging because superconductivity depends on complex compositional structure and subtle chemistry, making experimental exploration expensive and slow. Materials informatics offers a scalable alternative by learning relationships between chemical composition and critical temperature (Tc) from large datasets. Recent generative approaches have shown promise: GANs like ScGAN achieved 23-fold improvements over manual search rates, while diffusion models such as SuperDiff introduced conditioning on reference compounds to generate new superconductor families. However, these composition-level methods typically rely on vector encodings that do not explicitly model the sequential, symbolic structure of chemical formulas, which has proven effective for molecular generation. In this project, we develop a variational autoencoder (VAE) that learns to encode and decode superconductor chemical formulas while incorporating Tc and composition-derived material descriptors. Given a chemical formula together with Tc and engineered material features, the model produces an autoregressively reconstructed (and potentially novel) formula sequence, learning a continuous latent representation intended to support both accurate reconstruction and meaningful latent-space navigation.

Machine learning is well-suited to this task due to the availability of sufficient training data and because chemical formulas exhibit structured symbolic regularities that can be modeled effectively with sequence-based architectures. However, standard evaluation using random train/test splits can be misleading, since highly similar compositions may appear across splits and inflate performance without true chemical generalization. To address this, we adopt a generative evaluation strategy centered on whether the model can generate superconductors it has never observed during training. Our architecture uses a three-branch Full Materials Encoder that fuses learned element representations with Tc and 145 Magpie composition-derived features, which have shown strong predictive power for Tc in prior supervised learning studies. To support realistic stoichiometries, we introduce fraction-aware tokenization capable of representing both decimal and fractional coefficients. For decoding, we employ an autoregressive pointer–generator model with a copy mechanism for accurate formula reconstruction, augmented with learnable memory tokens to strengthen latent-to-sequence generation. This contrasts with prior superconductor generators that operate on fixed-dimensional composition vectors.

We train on the MDR SuperCon database containing approximately 32,000 superconductors with Tc values spanning 0–185 K and 145 Magpie features, with formulas represented in both decimal and fraction notation and family labels available for analysis, providing a comparable baseline to prior published models. We plan to expand beyond this dataset with additional sources such as OQMD for learning broader materials representations, NEMAD for contrastive learning between superconducting and magnetic behaviors, and SuperCon2 for improved coverage of high-pressure hydrides. Success is defined by both reconstruction performance and generative generalization: the model must achieve at least 90% exact-match formula reconstruction on training data, and it must generate previously unseen superconductors from a holdout set of 45 materials excluded entirely from training. We further test whether latent-space interpolation and sampling near high-Tc regions can recover withheld compounds (e.g., HgBa2Ca2Cu3O8, Tc ≈ 134 K), indicating the latent space captures superconducting chemistry rather than memorized strings. Overall, this sequence-based framework provides a scalable path toward generating chemically plausible superconductors while more rigorously evaluating generalization beyond training data.