

On the Importance of Descriptors in AI4Science

Krishna S Deb

February 27, 2026

1 What is a Descriptor

When we aim to predict properties of an object using ML/DL, we must first find a way to represent it using numerical values. For example, a table can be represented structurally as [number of legs, number of flat surfaces] or [length, breadth, height]. As shown in Figure 1, if we use [4,1] to represent a table, we cannot accurately predict its physical dimensions. Conversely, using only dimensions might not describe its structural stability (number of legs). Therefore, it is vital to study the input representation so the model can capture the necessary information. A descriptor is a numerical representation of an object used to describe it so that a model can map that information to a specific property.

2 Some common ‘objects’ of AI4Science

AI4Science spans various fields from physics and chemistry to biology and material science. Here, we discuss descriptors for molecules, magnetic materials, and particle physics.

2.1 Molecular Descriptors

In molecular representation learning, where the goal is to predict drug-related properties, **MoleculeNet** [13] is the gold standard. It utilizes domain-specific descriptors such as **ECFP** [8] and the **Coulomb Matrix** [10]. Modern approaches like **GROVER** [9], which uses Graph Neural Networks, represent molecules via atom and bond features. Atoms are described by features including atom type, formal charge, bond count, chirality, hydrogen count, atomic mass, aromaticity, and hybridization state (sp to sp^3d^2), forming a vector of size 128. Bond features include bond type, stereochemistry, and indicators for conjugation or ring membership.

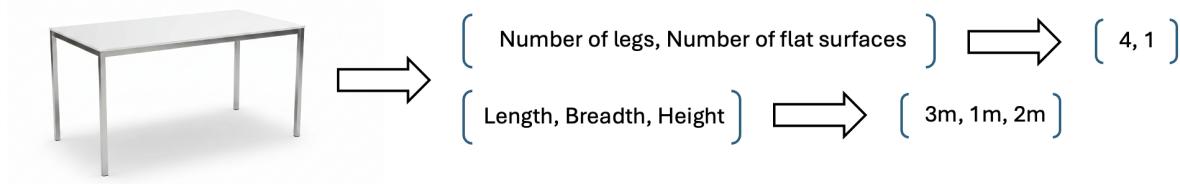


Figure 1: Example of descriptors of a Table

The **GeoGNN** model [2] further refines this by using a multi-level representation. It includes atom-level one-hot encodings for attributes like degree and hybridization, bond-level attributes including continuous bond length values, and geometric features such as the bond angle.

2.2 Magnetic Material Descriptors

The field of magnetic materials utilizes a diverse range of descriptors categorized by chemical, physical, and process-based attributes.

Elemental and Electronic Descriptors focus on intrinsic atomic properties. These include *f*-bandwidth, valence electrons, electronegativity, and atomic number [11]. Researchers also utilize atomic fractions weighted by elemental averages and deviations [11], as well as specific concentrations of metals (Fe, Ni, Co, Cr, Mn) and chalcogens [6]. Further descriptors include weight percentages (wt%) of rare-earth and transition elements [7], atomic percentages of chemical elements, Pauling electronegativity, and the work function of the material [14].

Thermodynamic and Physical Properties are critical for predicting performance under stress. These include molar volume, melting temperature, and bond length [11]. In dynamic environments, descriptors such as waveform, frequency (Hz), and magnetic flux density (T) are used [3]. Other features include theoretical density, atomic radius, and mixing entropy/enthalpy calculated via the rule of mixtures [14].

Structural and Process Descriptors capture the physical configuration and manufacturing history. Structural features include the location of substitutional sites [6], phase composition (ferromagnetic vs. impurity phases), and sample state—such as powder, sintered magnet, or ribbon [12]. Manufacturing parameters act as vital descriptors, including sintering and annealing temperatures/times [5, 7], as well as hot-extrusion temperatures and ram speeds [4].

2.3 Particle Physics Descriptors

In particle physics, descriptors (often called observables) are essential for both ML and Quantum ML (QML) pipelines. Based on [1], these are primarily categorized by energy and geometry. **Energy and Momenta** descriptors include Transverse Energy (E_T) for identifying energy deposition and Lepton/Bottom Jet Transverse Momenta (p_T) for reconstructing decay chains. **Angular Relationships** include the angle between leptons (θ_l) and the angular separation (ΔR_{l1}), defined via azimuthal angle and rapidity, which serve as indicators for particle isolation.

References

- [1] Callum Duffy, Mohammad Hassanshahi, Marcin Jastrzebski, and Sarah Malik. Unsupervised beyond-standard-model event discovery at the lhc with a novel quantum autoencoder. *Quantum Machine Intelligence*, 7(1):41, 2025.
- [2] Xiaomin Fang, Lihang Liu, Jieqiong Lei, Donglong He, Shanzhuo Zhang, Jingbo Zhou, Fan Wang, Hua Wu, and Haifeng Wang. Geometry-enhanced molecular representation learning for property prediction. *Nature Machine Intelligence*, 4(2):127–134, 2022.

- [3] Junqi He, Yifeng Wei, and Daiguang Jin. Modeling of core loss based on machine learning and deep learning. *arXiv preprint arXiv:2502.05487*, 2025.
- [4] Guillaume Lambard, TT Sasaki, K Sodeyama, T Ohkubo, and K Hono. Optimization of direct extrusion process for nd-fe-b magnets using active learning assisted by machine learning and bayesian optimization. *Scripta Materialia*, 209:114341, 2022.
- [5] Qichao Liang, Qiang Ma, Hao Wu, Rongshun Lai, Yangyang Zhang, Ping Liu, and Tao Qi. Performance prediction of sintered ndfeb magnet using multi-head attention regression models. *Scientific Reports*, 14(1):28822, 2024.
- [6] Dharmendra Pant, Suresh Pokharel, Subhasish Mandal, Dukka B Kc, and Ranjit Pati. Dft-aided machine learning-based discovery of magnetism in fe-based bimetallic chalcogenides. *Scientific Reports*, 13(1):3277, 2023.
- [7] Zuqiang Qiao, Shengzhi Dong, Qing Li, Xiangming Lu, Renjie Chen, Shuai Guo, Aru Yan, and Wei Li. Performance prediction models for sintered ndfeb using machine learning methods and interpretable studies. *Journal of Alloys and Compounds*, 963:171250, 2023.
- [8] David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of chemical information and modeling*, 50(5):742–754, 2010.
- [9] Yu Rong, Yatao Bian, Tingyang Xu, Weiyang Xie, Ying Wei, Wenbing Huang, and Junzhou Huang. Grover: Self-supervised message passing transformer on large-scale molecular data. *arXiv preprint arXiv:2007.02835*, 2(3):17, 2020.
- [10] Matthias Rupp, Alexandre Tkatchenko, Klaus-Robert Müller, and O Anatole Von Lilienfeld. Fast and accurate modeling of molecular atomization energies with machine learning. *Physical review letters*, 108(5):058301, 2012.
- [11] Prashant Singh, Tyler Del Rose, Andriy Palasyuk, and Yaroslav Mudryk. Physics-informed machine-learning prediction of curie temperatures and its promise for guiding the discovery of functional magnetic materials. *Chemistry of Materials*, 35(16):6304–6312, 2023.
- [12] Andrii Trostianchyn, Zoia Duriagina, Ivan Izonin, Roman Tkachenko, Volodymyr Kulyk, and Olena Pavliuk. Research paper sm-co alloys coercivity prediction using stacking heteroge-neous ensemble model. *Acta Metallurgica Slovaca*, 27(4):195–202, 2021.
- [13] Zhenqin Wu, Bharath Ramsundar, Evan N Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S Pappu, Karl Leswing, and Vijay Pande. Moleculenet: a benchmark for molecular machine learning. *Chemical science*, 9(2):513–530, 2018.
- [14] LI Xin, Chan Hung SHEK, et al. Domain knowledge aided machine learning method for properties prediction of soft magnetic metallic glasses. *Transactions of Nonferrous Metals Society of China*, 33(1):209–219, 2023.