MOFology: A Knowledge Graph for Engineering Direct Air Capture Materials

Anonymous Authors[1]

Abstract

Direct Air Capture (DAC) of CO_2 requires sorbents that are simultaneously CO_2-selective, water-tolerant, thermally regenerable, and synthesizable, with Metal-Organic Frameworks (MOFs) widely regarded as the most promising candidates. Yet the data describing MOFs is scattered across multiple databases in disparate formats, leaving cross-domain opportunities in DAC materials engineering unrealized. We present MOFology, an ontology-grounded knowledge graph (KG) that integrates $\sim$ 250,000 MOFs from six databases into $\sim$ 8.4 million RDF triples. The KG supports a suite of DAC-relevant tasks: chemically interpretable SPARQL semantic queries, graph embedding, concept vectors recovered as linear directions encoding chemical properties, prediction of change in CO_2 binding energy upon amine functionalization, and a multi-criteria DAC screen ranking over 9,000 real MOFs. To our knowledge, MOFology is the largest MOF KG to date and the first work to featurize the parent–derivative amine functionalization relation for predicting property changes, enabling a comprehensive database and prediction framework for DAC materials engineering.

1. Introduction

Direct Air Capture (DAC) of atmospheric CO_2 is a critical component of climate change mitigation efforts. (Sanz-Pérez et al., 2016) However, the requirements of sorbents to simultaneously display CO_2 selectivity, stability under humid conditions, thermal regenerability, and scalable synthesis represent a challenge to chemists, with few known materials adhering to all requirements at once. (Shi et al., 2020) Metal Organic Frameworks (MOFs) are a class of porous materials consisting of metal nodes held together by organic linkers, and have garnered attention for their diversity of chemical properties and myriad application directions. Thanks to their chemical tunability, MOFs are among the most promising DAC candidate classes. (Bose et al., 2024) At the same time, their chemical diversity means that there is a practically infinite number of possible MOFs, (Lee et al., 2021) with only a fraction of these materials having been explored experimentally or computationally.

Realizing the potential of MOFs for DAC is therefore limited not only by synthesis and characterization throughput, but by a combinatorial data problem. Furthermore, the known data on MOFs is fragmented across several data sources. A single MOF may be described by CSD (Moghadam et al., 2017) crystal codes, DFT-computed electronic properties, GCMC-simulated gas uptakes, experimental stability measurements, literature-extracted synthesis recipes, and post-synthetic modification records, each of which can be stored across different databases, under different identifiers, with different unit conventions. Simply concatenating these sources destroys the relationships that matter most for MOF reasoning, including which linker produced which pore geometry under which solvent, which functionalization shifted which binding energy in which parent framework, and which topology–metal combinations have actually been realized experimentally. As a result, conventional relational databases reduce this rich relational structure to flat tables suitable mainly for simple search and property regression. (An et al., 2022)

Knowledge graphs offer a natural abstraction for heterogeneous, relation-rich materials data. (Venugopal & Olivetti, 2024; Pruyn et al., 2025) By encoding MOFs as typed entities connected by typed relations, a knowledge graph preserves provenance, supports multi-hop semantic queries, and can be used to produce graph embeddings that downstream machine learning models can consume alongside standard chemical descriptors. (Fang et al., 2023) Prior MOF knowledge graphs, however, have focused predominantly on structural cataloging and are limited in their integration of DAC-specific property datasets, do not represent functionalization as a first-class relation, and have not been validated as substrates for predicting DAC-relevant proper-

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

Table 1. Comparison of MOFology with existing materials knowledge graphs.(An et al., 2022; Venugopal & Olivetti, 2024; Bai et al., 2025; Pruyn et al., 2025)

Work	Scope	# MOFs	# Triples/Rels.	Ontology Basis	Functionalization	DAC Focus
MOF-KG (An et al., 2022)	MOFs	∼125k	∼3.7M rels.	Custom (Neo4j)	X	X
MatKG (Venugopal & Olivetti, 2024)	All materials	N/A (70k entities)	5.4M triples	Statistical co-occ.	X	X
KG-FM (Bai et al., 2025)	MOFs/COFs/HOFs	Not MOF-specific	4.01M rels.	LLM-extracted	X	X
MOF-ChemUnity (Pruyn et al., 2025)	MOFs	∼15k	∼3.2M rels.	Custom (Neo4j)	X	X
MOFology (this work)	**MOFs**	**∼254k**	**8.4M triples**	**EMMO (OWL)**	✓	✓

ties. (Venugopal & Olivetti, 2024; Pruyn et al., 2025)

Here we present MOFology, a knowledge graph designed explicitly for DAC-oriented MOF engineering. We integrate 250,000 MOFs from six databases into 8.4 million RDF triples under an EMMO (Del Nostro et al., 2024)-aligned ontology that includes first-class representation of amine functionalization and parent–derivative relations. We demonstrate MOFology as (i) a semantic search engine supporting chemist-level SPARQL queries; (ii) a substrate for graph representation learning, benchmarking three embedding methods (CompGCN (Vashishth et al., 2019), Node2Vec (Grover & Leskovec, 2016), TransE (Bordes et al., 2013)) under family-aware evaluation; (iii) a tool for mechanistic interpretation, showing that linear probes recover chemically meaningful concept vectors from the embedding space, and (iv) a platform for multi-criteria DAC screening that ranks a subset of 9,000 experimentally reported MOFs whose DAC properties were absent from the KG. Table 1 compares the coverage and purpose of MOFology to other materials science knowledge graphs.

2. Methods

2.1. Data Collection and Curation

MOFology integrates six databases covering MOF chemistry from physicochemical properties to synthesis procedures. OpenDAC25 (Sriram et al., 2025) contributes CO_2 and H_2O binding energies and amine-functionalization pairs, DigiMOF (Glasby et al., 2023) supplies literature-extracted synthesis conditions and linker provenance, QMOF (Rosen et al., 2021) provides DFT-level electronic properties, MOF-ChemUnity (Pruyn et al., 2025) contributes unified experimental properties, MOF-FreeEnergy (Niyongabo Rubungo et al., 2025) provides thermodynamic stabilities, and SynMOF (Luo et al., 2022) contributes synthesis information. MOF identifiers were harmonized across databases using CSD reference codes, Materials Project identifiers, and MOFid (Bucior et al., 2019) structural descriptors as cross-reference keys. Data curation followed established best practices in materials informatics (Hart et al., 2024). Chemical formulas were standardized using the pymatgen (Ong et al., 2013) composition parser and SMILES strings for organic linkers were canonicalized using RDKit. Table 7 in the appendix summarizes the per-source MOF counts and property-family coverage.

2.2. Ontology Development

The MOFology ontology acts as the basis for the construction of the knowledge graph, offering entity and relationship schemas and connection rules to ensure the KG remains interoperable and maintainable. We extend the Elementary Multiperspective Material Ontology (EMMO) (Del Nostro et al., 2024) with domain-specific classes designed to capture the entities and relations relevant to DAC engineering. A `MOF` branch is introduced containing `ExperimentalMOF` for structures that have been experimentally synthesized, `HypotheticalMOF` for structures that have been computationally studied, but contain no evidence for having been experimentally synthesized, and `FunctionalizedMOF` for MOFs that have been modified with a post-synthetic functionalization. The `Material` branch is extended with `MetalCluster` and `Linker` acting as the domain from which MOFs are constructed. The `Process` branch consists of `Synthesis` and `Functionalization`, and the `Property` branch divides all properties associated with MOFs into `Structural`, `Computational`, and `Experimental`. Figure 1 summarizes the class hierarchy with an example instantiation.

After manual axiom authoring, we use the HermiT reasoner (Glimm et al., 2014) to enforce ontological consistency and materialize implicit knowledge. The reasoner checks class-disjointness constraints, ensuring no MOF is simultaneously classified as both experimental and hypothetical, and verifies domain and range restrictions on all object properties. The reasoner also materializes inverse-property axioms.

2.3. KG Construction

Source-specific extractors were constructed in Python to transform each database into entity and predicate instances that conform to the ontology. After base-graph instantiation, an OWL reasoner (HermiT) (Glimm et al., 2014) applies inverse-property, transitive, and property-chain inference, materializing MOF classifications and parent–derivative relationships that would otherwise require query-time expansion. The final graph comprises 8.4M triples serialized in TTL format. A complete breakdown of the knowledge graph

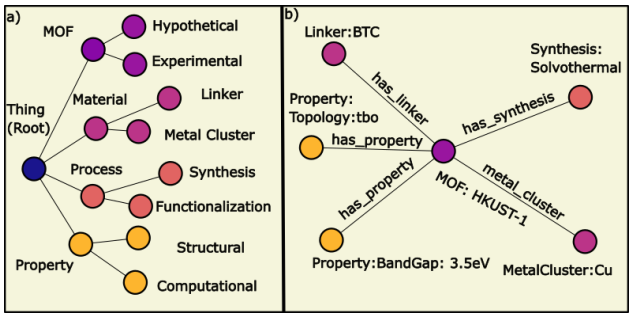

Figure 1. MOFology ontology extending EMMO with MOF-specific classes covering materials, properties, synthesis, and functionalization. a) An overview of the classes in the Ontology. b) What an instantiation of the MOF HKUST-1 (Britt et al., 2008) looks like in this ontological framework.

statistics can be found in Table 2.

2.4. Embedding Methods and Learning Tasks

We train three 256-dimensional KG embeddings: TransE as a naive baseline, Node2Vec for learning graph topology, and CompGCN for learning node and edge features and topology simultaneously. TransE (Bordes et al., 2013) models each relation as a translation in embedding space. For an observed triple (h, r, t), TransE learns embeddings $\mathbf{h}, \mathbf{r}, \mathbf{t} \in \mathbb{R}^d$ such that $\mathbf{h}+\mathbf{r} \approx \mathbf{t}$, optimizing a margin-based ranking loss:

$$\mathcal{L}_{\text{TransE}} = \sum_{(h,r,t)\in\mathcal{T}} \sum_{(h',r,t')\in\mathcal{T}'} \Big[\gamma + d(\mathbf{h}+\mathbf{r},\, \mathbf{t}) - d(\mathbf{h}'+\mathbf{r},\, \mathbf{t}')\Big]_+ \tag{1}$$

where $\mathcal{T}$ is the set of observed triples, $\mathcal{T}'$ is a set of corrupted triples generated by replacing either the head or tail entity, γ is the margin hyperparameter, $d(\cdot,\cdot)$ is the L_2 distance, and $[\cdot]_+$ denotes the positive part.

Node2Vec (Grover & Leskovec, 2016) generates biased random walks over the graph controlled by a return parameter p and an in-out parameter q, then trains a skip-gram model to maximize:

$$\mathcal{L}_{\text{N2V}} = \sum_{v\in V} \sum_{u\in\mathcal{C}(v)} \Big[-\log\sigma\big(\mathbf{h}_u^\top \mathbf{h}_v\big) - \sum_{i=1}^{k} \mathbb{E}_{v_i\sim P_n} \log\sigma\big(-\mathbf{h}_u^\top \mathbf{h}_{v_i}\big)\Big] \tag{2}$$

where $\mathcal{C}(v)$ is the set of context entities co-occurring with v in random walks and σ is the sigmoid function. The walk bias interpolates between depth-first (high p, low q) and breadth-first (low p, high q) exploration, allowing Node2Vec to capture local neighborhood structure while encoding the global topology of the KG.

CompGCN (Vashishth et al., 2019) is a relation-composition graph convolutional network that jointly learns entity and relation representations through message passing. At each layer ℓ, entity representations are updated by aggregating relation-typed messages from neighbors:

$$\mathbf{h}_v^{(\ell+1)} = f\Bigg(\sum_{(u,r)\in\mathcal{N}(v)} \mathbf{W}_O^{(\ell)}\, \phi\big(\mathbf{h}_u^{(\ell)}, \mathbf{h}_r^{(\ell)}\big) + \sum_{(u,r)\in\mathcal{N}^{-1}(v)} \mathbf{W}_I^{(\ell)}\, \phi\big(\mathbf{h}_u^{(\ell)}, \mathbf{h}_{r^{-1}}^{(\ell)}\big) + \mathbf{W}_S^{(\ell)}\, \mathbf{h}_v^{(\ell)} \Bigg) \tag{3}$$

where $\mathcal{N}(v)$ denotes the typed neighborhood of entity v, ϕ is a composition operator (we use subtraction), $\mathbf{W}_{\lambda(r)}^{(\ell)}$ is a direction-specific weight matrix with $\lambda(r) \in \{\text{original}, \text{inverse}, \text{self-loop}\}$, and f is a nonlinear activation. We use a 3-layer architecture.

As a non-graph baseline, we compute standard chemical features: RDKit (Landrum et al., 2013) Morgan fingerprints (radius 2, 2048-bit) on canonical linker SMILES and pymatgen (Ong et al., 2013) composition descriptors on the metal cluster formula. A hybrid feature set concatenates chemical features with KG embeddings.

We evaluate the KG and embeddings on DAC-motivated tasks: property prediction via regression of physicochemical properties across three previously mentioned feature families (KG embeddings, Chemical features, and Hybrid) using family-aware train/test splits that prevent parent–derivative leakage; link prediction via binary classification of edge existence using operator features over source and target embeddings with negative sampling; functionalization prediction via regression of change of CO_2 and H_2O binding energy between parent and amine-functionalized MOFs using embedding differences and chemical deltas; and concept vector (Graziani et al., 2023) probing, where for each chemically meaningful concept c (e.g., "high density," "topology = pcu," "metal = Zn"), we train a logistic regression classifier on standardized embeddings to separate MOFs exhibiting the concept from those that do not. For continuous properties, MOFs above the 75th percentile are labeled positive and those below the 25th percentile are labeled negative; for categorical labels, a one-vs-all scheme is used. The unit-normalized weight vector $\hat{\mathbf{w}}_c = \mathbf{w}_c/\|\mathbf{w}_c\|$ defines a *concept direction* in embedding space. Projecting any MOF embedding $\mathbf{h}$ onto this direction,

$$s_c = \mathbf{h}\cdot\hat{\mathbf{w}}_c\,, \tag{4}$$

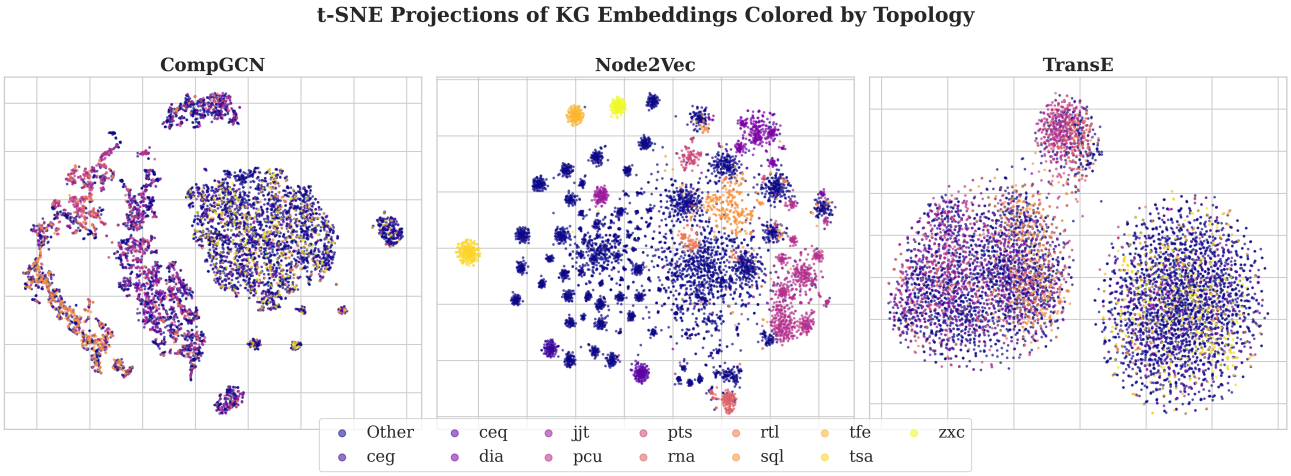

Figure 2. Three Embedding methods were used in property prediction experiments and KG tasks. Here is shown the t-SNE structure of CompGCN, Node2Vec, and TransE, colored by MOF topology. CompGCN shows a more organized continuous structure than the other examples, but Node2Vec is better able to separate MOF embeddings by topology. TransE was used as a naive baseline for embedding studies.

Table 2. MOFology Knowledge Graph statistics.

Statistic	Count
Total MOFs	254,099
Experimental (CSD)	40,791
Hypothetical	213,308
Functionalized (amine-grafted)	2,650
Total Triples	8,399,032
Unique Organic Linkers	∼5,000
Unique Metal Clusters	∼1,400
Unique Topologies	∼1,000
Distinct Properties	95
Properties (≥500 samples)	49
Relation Types	15
Data Sources	6

yields a scalar score for how strongly that MOF exhibits concept c.

3. Results and Discussion

3.1. Knowledge Graph Overview

MOFology integrates 250,000 MOFs into 8.4M RDF triples: 40,791 experimental and 213,308 hypothetical MOFs. Table 2 summarizes statistics.

3.2. Semantic Searching

One of the principal advantages that KGs display over traditional relational databases is multi-hop semantic searching in SPARQL statements. MOFology supports chemist-level semantic queries that would require manual joins across multiple databases in a tabular setting. Table 3 demonstrates four such queries over the full graph. These include filtering experimental MOFs by pore size and density (847 matches), identifying MOFs with strong CO_2 binding but weaker H_2O affinity by joining binding energies across sources (312 matches), retrieving Mg-containing MOFs of specific topologies (23 matches), and traversing derivedFrom relations to find the 1,892 amine-functionalized derivatives whose CO_2 binding improved over their parent. Each query executes in under one second on the SPARQL endpoint.

3.3. Embeddings

We benchmark three embedding methods under family-aware evaluation that prevents parent–derivative leakage (Table 4). Node2Vec achieves highest link prediction (AUC $= 0.976$); CompGCN achieves highest average property prediction performance ($R^2 = 0.467$), indicating that its relation-typed message passing learns representations more aligned with physicochemical properties than Node2Vec's topology-focused embeddings. TransE, included as a baseline, underperforms on both tasks, likely because the single-vector-per-relation assumption is too restrictive for the heterogeneous relation types in MOFology. Both CompGCN and a combination of CompGCN and chemical features either beat or are competitive with chemical features alone for several properties. Figure 8 in the appendix shows property prediction R^2 by embedding and feature family. Hybrid features (KG embeddings concatenated with chemical descriptors) achieve $R^2 = 0.63$ for CompGCN and $R^2 = 0.63$ for Node2Vec, competitive with the ChemOnly baseline ($R^2 = 0.66$). This near-parity is notable: the KG embeddings encode chemical information learned purely from graph structure, without any explicit molecular featurization, and yet improve the predictive signal in several instances (Figure 8).

Table 3. DAC-chemist SPARQL queries over the MOFology KG .

Chemist Question	Query Pattern	Result summary
Experimental MOFs with large pores and low density (DAC-capacity candidates).	PLD $>$ 6 Å $\wedge$ Density $<$ 1.0 g/cm^3	847 matches. E.g. ABEXEN (PLD=13.7 Å, ρ=0.67); ACOLEL (PLD=10.1 Å, ρ=0.29).
Strong CO_2 binders with weaker H_2O affinity.	CO_2 BE $<$ -0.5 eV $\wedge$ CO_2 BE $<$ H_2O BE	312 matches. E.g. CUGVUW (CO_2=-1.14, H_2O=-0.82 eV); LUSYEF (CO_2=-0.62, H_2O=-0.45 eV).
Mg-containing MOFs with pcu/fcu topology (Mg-MOF-74 analogues).	hasMetalElement = Mg $\wedge$ topologyCode $\in$ {pcu, fcu}	23 matches. E.g. COKPEZ (pcu); COQTEJ (pcu); DOHDOU (pcu).
Amine derivatives whose functionalization improves CO_2 binding vs. their parent.	syn:derivedFrom $\wedge$ child CO_2 BE $<$ parent CO_2 BE	1,892 matches. E.g. FuncMOF_ABEXEM $\rightarrow$ ABEXEM; FuncMOF_ABUWOJ $\rightarrow$ ABUWOJ.

The t-SNE projections in Figure 2 visualize these complementary strengths. CompGCN embeddings exhibit a smoother, more continuous structure in which MOFs of similar properties occupy nearby regions, while Node2Vec produces tighter, more clearly separated clusters organized by topology. TransE embeddings show less interpretable structure, consistent with their lower downstream performance.

KG embeddings also enable label imputation for hypothetical MOFs lacking chemical descriptors (Table 5). Node2Vec achieves 0.78 accuracy on topology classification using only graph structure. The two methods capture complementary aspects of the graph: CompGCN learns property-relevant representations through relation-typed message passing, while Node2Vec preserves local neighborhood structure that encodes topology and metal identity.

3.4. Concept Vectors

Linear probes trained on CompGCN embeddings, as described in section 2, reveal that chemically meaningful properties are encoded as near-linear directions in the embedding space. Property-based concepts such as density, pore limiting diameter, band gap, and unit cell volume, achieve ROC-AUC above 0.95, indicating that these properties are recoverable from embeddings with high fidelity (Figure 3). Importantly, high and low directions for the same property (e.g., "high unit cell volume" and "low unit cell volume") emerge as approximate inverses. Figure 9 in the Appendix displays a heatmap of positively and negatively correlated concepts. This indicates that continuous properties are encoded along oriented axes that enables the amine functionalization prediction below.

3.5. Functionalization Prediction

Predicting the change in binding energy upon amine grafting is a task enabled by MOFology's `derivedFrom` relations, which pair each functionalized MOF with its unmodified parent. Using CompGCN embedding differences ($\mathbf{h}_{m'} - \mathbf{h}_m$) as features, a linear probe achieves $R^2 = 0.893$ for ΔCO_2 binding energy (Figure 4), indicating that MOFology can reliably rank functionalization candidates by their predicted CO_2 affinity gain. This result validates the concept vector analysis section. The embedding difference between a parent and its amine-grafted derivative aligns with the learned concept direction, and projection magnitude correlates with the magnitude of the binding energy shift. ΔH_2O binding is substantially harder to predict ($R^2 = 0.247$), with larger residuals and only a mild trend in the error structure (Figure 4). This gap likely reflects two factors: first, water binding in MOFs involves extended hydrogen-bonding networks that are not captured by local coordination descriptors or pairwise binding energies; second, H_2O binding data in the current graph is sparse and derives almost entirely from a single source, limiting the diversity of training examples. Improving ΔH_2O prediction will require both richer water-interaction descriptors and additional data ingestion from emerging experimental and computational sources.

3.6. Multi-Criteria DAC Screening

Using the hybrid models derived from the CompGCN embeddings and chemical features, we screen approximately 9,000 real MOFs combining CO_2 affinity, hydrophobicity, and stability with reliability weights (Figure 5), which were present in the KG but lacked all explicit properties required for DAC assessment. Table 8 details how each predicted property contributes to the overall DAC score. The top candidate is MOF HUZFOY (score = 0.84, Figure 6), a pcu-topology MOF. Table 6 lists the top five. Four of the top five candidates adopt pcu topology, consistent with the

Table 4. KG embedding benchmark. Link prediction (LP) is reported as mean ± SD over 3 seeds on family-aware holdouts that prevent parent–derivative leakage. Property prediction is reported as mean best R^2 ± SD across several targets. TransE included as a single-vector-per-relation baseline. Its underperformance shows MOFology's heterogeneous relations require more sophisticated embedding methods.

METHOD	LP AUC	LP HITS@10	KG-ONLY R^2	HYBRID R^2
COMPGCN	0.895 ± 0.005	0.907 ± 0.006	0.467 ± 0.362	0.626 ± 0.359
NODE2VEC	0.976 ± 0.002	0.995 ± 0.002	0.243 ± 0.171	0.627 ± 0.307
TRANSE	0.884 ± 0.001	0.808 ± 0.008	0.066 ± 0.089	0.595 ± 0.337
CHEMONLY (BASELINE)	—	—	—	0.660 ± 0.316

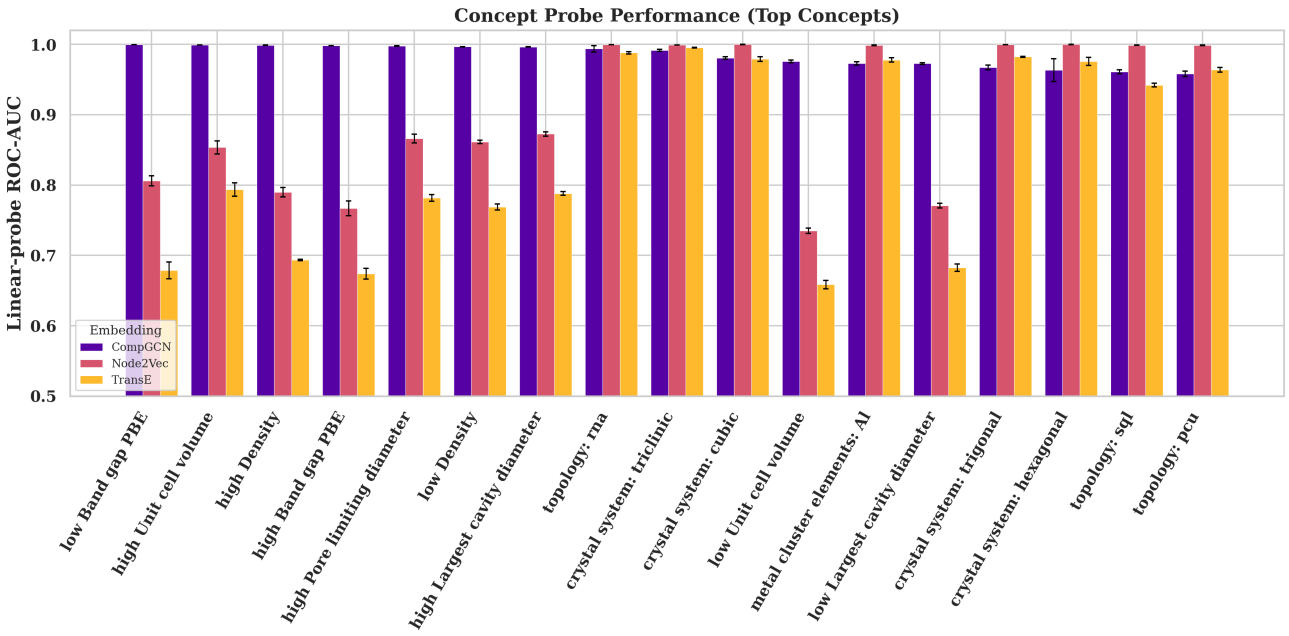

Figure 3. Concept vector probing according to each embedding method.

Table 5. Label imputation from KG embeddings. Accuracy and weighted F1 for predicting topology and metal element using only KG embeddings (no chemical features). Methods are compared to most frequent class baseline.

TARGET	EMBEDDING	ACC.	WTD. F1	BASELINE
TOPOLOGY	NODE2VEC	0.782	0.749	0.100
	COMPGCN	0.263	0.219	0.100
	TRANSE	0.234	0.160	0.100
METAL	NODE2VEC	0.685	0.633	0.188
	COMPGCN	0.427	0.371	0.188
	TRANSE	0.355	0.252	0.188

Table 6. Top DAC candidate MOFs and their scores from the multi-criteria screen.

MOF	TOPOLOGY	DAC
MOF_HUZFOY	PCU	0.838
MOF_LUCDAO	PCU	0.789
MOF_OJAKOA	PCU	0.778
MOF_ODAHIK	CDS	0.758
MOF_YARSIV	PCU	0.738

known suitability of pcu frameworks for gas capture applications. (Anderson et al., 2018; Shabangu et al., 2025) These rankings are predictions from the KG-derived scoring pipeline and await experimental and DFT-level validation. When the same screening campaign was performed on MOF candidates for which the DAC properties and applications were well-known, top DAC candidates included MOF COKPEZ, which has been extensively studied for carbon capture. (Queen et al., 2014)

The top-ranked candidate from the multi-criteria screen, a MOF with the CSD code HUZFOY, illustrates the type of non-obvious candidate the KG framework can surface. HUZFOY is a cadmium-based pcu coordination complex constructed from rigid bis(benzimidazolyl) ligands, first reported in ref (Li et al., 2010) in a study focused on tuning porosity and interpenetration through ligand modification

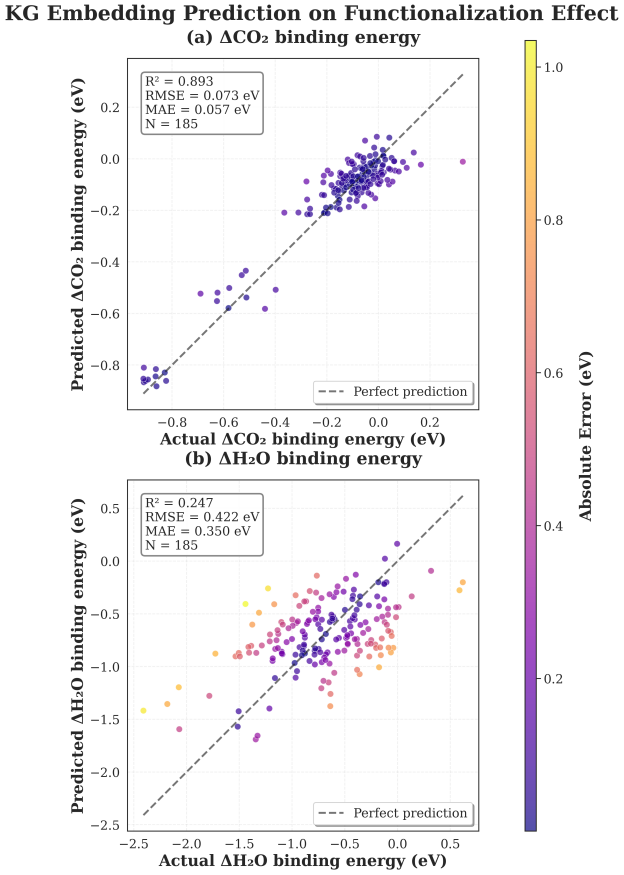

Figure 4. Functionalization prediction using uncovered concept vectors from KG embeddings. ΔCO_2 is well-predicted across feature families. ΔH_2O binding signal was weak in the KG.

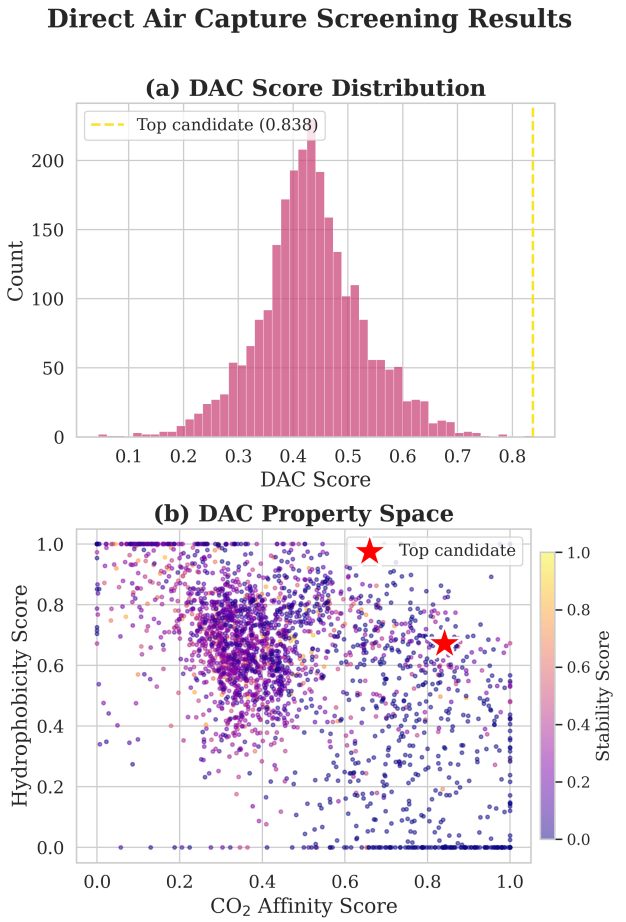

Figure 5. DAC screening over 9,085 MOFs. (a) Composite score distribution. (b) CO_2 affinity vs. hydrophobicity colored by stability. Top candidate (starred, MOF HUZFOY) simultaneously displays CO_2-affinity, hydrophobicity, and stability.

(Figure 6). The original work characterized the structural and adsorption properties of the framework but did not assess its capture performance. To our knowledge, HUZFOY has not been subsequently evaluated for carbon capture in the literature. Its emergence as the top DAC candidate is driven entirely by KG-derived features, including its pcu topology, pore geometry, computed thermodynamic stability, and embedding-space proximity to known high-performing sorbents. This prediction awaits experimental validation through CO2 and H2O adsorption measurements under DAC-relevant conditions, and represents a concrete, testable hypothesis generated by the MOFology framework.

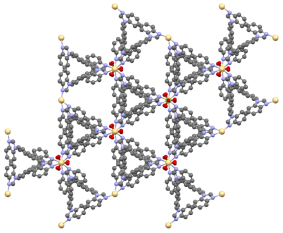

Figure 6. DAC screening revealed the MOF with the CSD code HUZFOY, first described in reference (Li et al., 2010), as the best candidate for carbon capture applications.

4. Conclusion

MOFology enables a DAC-engineering workflow that traditional tabular databases cannot support. A researcher can narrow a search space of over 250,000 MOFs using semantic queries that jointly filter on metal identity, topology, property coverage, and synthesis history as well as rank the resulting candidates using reliability-weighted multi-criteria scoring that balances CO_2 affinity, hydrophobicity, and stability. Additionally, for candidates lacking experimental binding data, it can predict the change in CO_2 binding energy upon amine functionalization using concept vectors learned directly from the graph. This workflow moves from broad exploration to targeted prediction without leaving the KG, eliminating the manual cross-referencing across disparate databases that currently slows MOF-based sorbent development.

The KG abstraction earns its added complexity over simpler tabular representations. CompGCN embeddings achieve reasonable prediction performance for property prediction using only graph-derived features, without any explicit chemical descriptors, demonstrating that the relational semantics of the graph encode chemistry-relevant structure rather than simply storing it. The near-parity between hybrid and chemistry-only baselines further confirms that graph connectivity captures much of the same information as molecular fingerprints and composition descriptors. Concept vector probing reveals that this information is organized along interpretable linear directions in embedding space, and the amine-functionalization direction translates directly to quantitative prediction of CO_2 binding energy changes, connecting representation learning to a physically actionable quantity.

Several limitations motivate future work. First, while the KG-derived embeddings produce features that compete with chemical descriptors in the modeling of several properties, the average performance of the ChemOnly models still surpasses that of the KG and hybrid features. This indicates that the strengths of the KG lie in concept vector analysis, semantic search, link prediction, and data imputation. Second, DAC-specific property coverage remains sparse: CO_2 and H_2O binding energies are available for only approximately 2,600 MOFs, drawn almost entirely from OpenDAC25. This sparsity constrains both the multi-criteria screening and the functionalization prediction, particularly for water binding, where the current graph under-predicts ΔH_2O due to the absence of features capturing extended hydrogen-bonding networks. As new DAC-relevant datasets emerge from ongoing experimental and computational campaigns, the ontology schema can absorb them without structural modification, and we expect prediction quality to improve accordingly. The top-ranked DAC candidates identified by the screening pipeline are predictions from KG-derived scores and await experimental synthesis and characterization under realistic DAC conditions. Validating these predictions through targeted experiments would close the loop between computational screening and laboratory realization.

More broadly, the approach demonstrated here is not specific to DAC or to MOFs. The combination of ontology-grounded knowledge graph construction, family-aware embedding evaluation, and concept vector probing could be applied to any materials class where heterogeneous data sources, compositional diversity, and post-synthetic modification create the same kind of relational data problem. MOFology is, to our knowledge, the largest MOF knowledge graph to date and the first to encode functionalization as a first-class relation. The KG, ontology, and embedding pipeline provide a reproducible foundation for DAC-focused MOF discovery and a template for knowledge-graph-driven materials engineering beyond the MOF domain.

Software and Data

The code and data will be released under an open source license upon acceptance.

Acknowledgements

This section is intentionally left blank for review.

Impact Statement

This paper presents a knowledge graph framework for accelerating the discovery of Metal-Organic Framework sorbents for Direct Air Capture of CO_2. By enabling efficient multi-criteria screening and property prediction for DAC materials, this work has the potential to accelerate the development of scalable carbon capture solutions, though the identified candidate materials require experimental validation before real-world deployment.

References

An, Y., Greenberg, J., Zhao, X., Hu, X., McCLellan, S., Kalinowski, A., Uribe-Romo, F. J., Langlois, K., Furst, J., Gómez-Gualdrón, D. A., et al. Building open knowledge graph for metal-organic frameworks (mof-kg): Challenges and case studies. *arXiv preprint arXiv:2207.04502*, 2022.

Anderson, R., Rodgers, J., Argueta, E., Biong, A., and Gómez-Gualdrón, D. A. Role of pore chemistry and topology in the co2 capture capabilities of mofs: from molecular simulation to machine learning. *Chemistry of Materials*, 30(18):6325–6337, 2018.

Bai, X., He, S., Li, Y., Xie, Y., Zhang, X., Du, W., and Li, J.-R. Construction of a knowledge graph for framework material enabled by large language models and its application. *npj Computational Materials*, 11(1):51, 2025.

Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., and Yakhnenko, O. Translating embeddings for modeling multi-relational data. *Advances in neural information processing systems*, 26, 2013.

Bose, S., Sengupta, D., Rayder, T. M., Wang, X., Kirlikovali, K. O., Sekizkardes, A. K., Islamoglu, T., and Farha, O. K. Challenges and opportunities: metal–organic frameworks for direct air capture. *Advanced Functional Materials*, 34 (43):2307478, 2024.

Britt, D., Tranchemontagne, D., and Yaghi, O. M. Metal-organic frameworks with high capacity and selectivity for harmful gases. *Proceedings of the National Academy of Sciences*, 105(33):11623–11627, 2008.

Bucior, B. J., Rosen, A. S., Haranczyk, M., Yao, Z., Ziebel, M. E., Farha, O. K., Hupp, J. T., Siepmann, J. I., Aspuru-Guzik, A., and Snurr, R. Q. Identification schemes for metal–organic frameworks to enable rapid search and cheminformatics analysis. *Crystal Growth & Design*, 19 (11):6682–6697, 2019.

Del Nostro, P., Friis, J., Ghedini, E., Goldbeck, G., Holtz, O., Roscioni, O. M., Zaccarini, F. A., Toti, D., et al. Elementary multiperspective material ontology: Leveraging perspectives via a showcase of emmo-based domain and application ontologies. *IC3K*, 2:135–142, 2024.

Fang, Y., Zhang, Q., Zhang, N., Chen, Z., Zhuang, X., Shao, X., Fan, X., and Chen, H. Knowledge graph-enhanced molecular contrastive learning with functional prompt. *Nature Machine Intelligence*, 5(5):542–553, 2023.

Glasby, L. T., Gubsch, K., Bence, R., Oktavian, R., Isoko, K., Moosavi, S. M., Cordiner, J. L., Cole, J. C., and Moghadam, P. Z. Digimof: a database of metal–organic framework synthesis information generated via text mining. *Chemistry of Materials*, 35(11):4510–4524, 2023.

Glimm, B., Horrocks, I., Motik, B., Stoilos, G., and Wang, Z. Hermit: an owl 2 reasoner. *Journal of automated reasoning*, 53(3):245–269, 2014.

Graziani, M., Mahony, L. O., Nguyen, A.-p., Müller, H., and Andrearczyk, V. Uncovering unique concept vectors through latent space decomposition. *arXiv preprint arXiv:2307.06913*, 2023.

Grover, A. and Leskovec, J. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 855–864, 2016.

Hart, M., Idanwekhai, K., Alves, V. M., Miller, A. J., Dempsey, J. L., Cahoon, J. F., Chen, C.-H., Winkler, D. A., Muratov, E. N., and Tropsha, A. Trust not verify? the critical need for data curation standards in materials informatics. *Chemistry of Materials*, 36(19):9046–9055, 2024.

Landrum, G. et al. Rdkit documentation. *Release*, 1(1-79): 4, 2013.

Lee, S., Kim, B., Cho, H., Lee, H., Lee, S. Y., Cho, E. S., and Kim, J. Computational screening of trillions of metal–organic frameworks for high-performance methane storage. *ACS Applied Materials & Interfaces*, 13(20):23647–23654, 2021.

Li, Z.-X., Hu, T.-L., Ma, H., Zeng, Y.-F., Li, C.-J., Tong, M.-L., and Bu, X.-H. Adjusting the porosity and interpénétration of cadmium (ii) coordination polymers by ligand modification: syntheses, structures, and adsorption properties. *Crystal growth & design*, 10(3):1138–1144, 2010.

Luo, Y., Bag, S., Zaremba, O., Cierpka, A., Andreo, J., Wuttke, S., Friederich, P., and Tsotsalas, M. Mof synthesis prediction enabled by automatic data mining and machine learning. *Angewandte Chemie International Edition*, 61 (19):e202200242, 2022.

Moghadam, P. Z., Li, A., Wiggin, S. B., Tao, A., Maloney, A. G., Wood, P. A., Ward, S. C., and Fairen-Jimenez, D. Development of a cambridge structural database subset: a collection of metal–organic frameworks for past, present, and future. *Chemistry of materials*, 29(7):2618–2625, 2017.

Niyongabo Rubungo, A., Fajardo-Rojas, F., Gómez-Gualdrón, D. A., and Dieng, A. B. Highly accurate and fast prediction of mof free energy via machine learning. *Journal of the American Chemical Society*, 147(52): 48035–48045, 2025.

Ong, S. P., Richards, W. D., Jain, A., Hautier, G., Kocher, M., Cholia, S., Gunter, D., Chevrier, V. L., Persson, K. A., and Ceder, G. Python materials genomics (pymatgen): A robust, open-source python library for materials analysis. *Computational Materials Science*, 68:314–319, 2013.

Pruyn, T. M., Aswad, A., Khan, S. T., Huang, J., Black, R., and Moosavi, S. M. Mof-chemunity: Literature-informed large language models for metal–organic framework research. *Journal of the American Chemical Society*, 147 (47):43474–43486, 2025.

Queen, W. L., Hudson, M. R., Bloch, E. D., Mason, J. A., Gonzalez, M. I., Lee, J. S., Gygi, D., Howe, J. D., Lee, K., Darwish, T. A., et al. Comprehensive study of carbon dioxide adsorption in the metal–organic frameworks m 2 (dobdc)(m= mg, mn, fe, co, ni, cu, zn). *Chemical Science*, 5(12):4569–4581, 2014.

Rosen, A. S., Iyer, S. M., Ray, D., Yao, Z., Aspuru-Guzik, A., Gagliardi, L., Notestein, J. M., and Snurr, R. Q. Machine learning the quantum-chemical properties of metal–organic frameworks for accelerated materials discovery. *Matter*, 4(5):1578–1597, 2021.

Sanz-Pérez, E. S., Murdock, C. R., Didas, S. A., and Jones, C. W. Direct capture of co2 from ambient air. *Chemical reviews*, 116(19):11840–11876, 2016.

Shabangu, S. M., Eaby, A. C., Nikkhah, S. J., Croitor, L., He, T., Bezrukov, A. A., Vandichel, M., and Zaworotko, M. J. A pcu topology metal–organic framework, ni (1, 4-bib)(inca) 2, that exhibits high co 2/n 2 selectivity and low water vapour affinity. *Journal of Materials Chemistry A*, 13(23):17562–17568, 2025.

Shi, X., Xiao, H., Azarabadi, H., Song, J., Wu, X., Chen, X., and Lackner, K. S. Sorbents for the direct capture of co2 from ambient air. *Angewandte Chemie International Edition*, 59(18):6984–7006, 2020.

Sriram, A., Brabson, L. M., Yu, X., Choi, S., Abdelmaqsoud, K., Moubarak, E., de Haan, P., Löwe, S., Brehmer, J., Kitchin, J. R., et al. The open dac 2025 dataset for sorbent discovery in direct air capture. *arXiv preprint arXiv:2508.03162*, 2025.

Vashishth, S., Sanyal, S., Nitin, V., and Talukdar, P. Composition-based multi-relational graph convolutional networks. *arXiv preprint arXiv:1911.03082*, 2019.

Venugopal, V. and Olivetti, E. Matkg: An autonomously generated knowledge graph in material science. *Scientific Data*, 11(1):217, 2024.

Appendix

A. Knowledge Graph Data Sources

Table 7. Data sources integrated into MOFology.

Source	MOFs	Key Properties
QMOF	20,375	DFT band gaps, pore sizes, density
OpenDAC25	2,237	CO_2/H_2O binding, amine functionalization
DigiMOF	15,143	Literature-extracted synthesis conditions
ChemUnity	2,053	Water stability, applications
MOF-FreeEnergy	213,308	Strain energy, free energy
SynMOF	983	Synthesis Information

B. Property Coverage

Figure 7 shows the distribution of property coverage across 253,184 MOFs in MOFology.

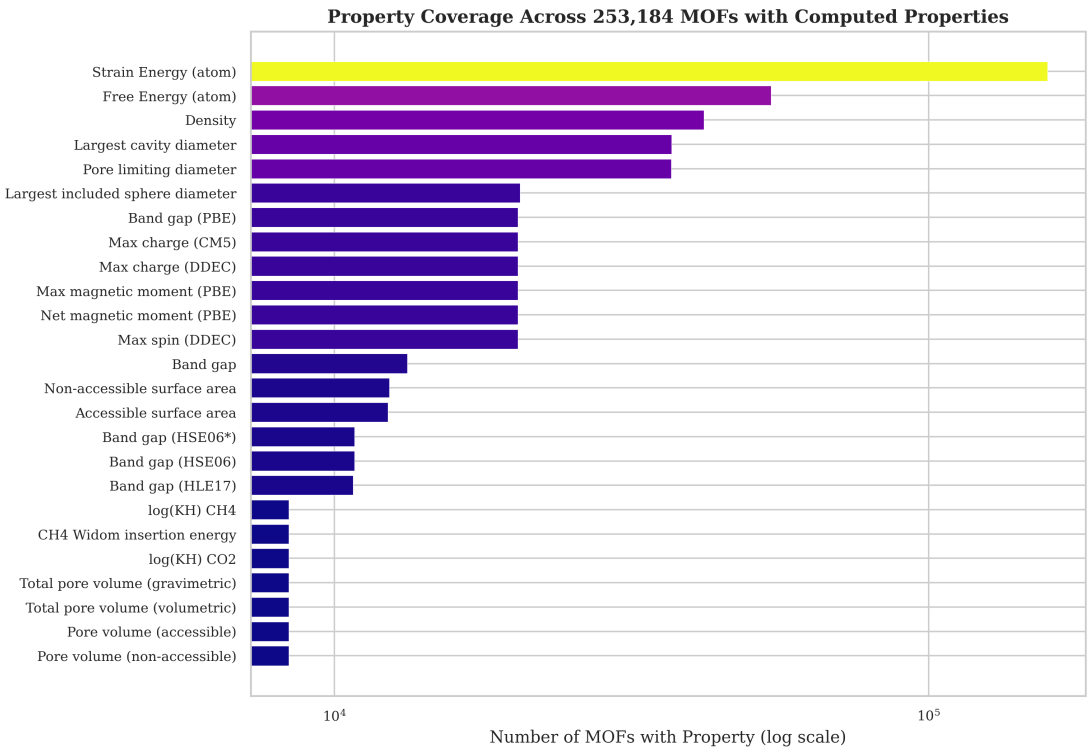

Figure 7. Property coverage across MOFs with computed properties. Thermodynamic properties from MOF-FreeEnergy dominate; DAC-relevant binding energies are available for ∼2,500 MOFs.

C. Property Prediction Details

Figure 8 shows the full R^2 matrix across embeddings, feature families, and target properties.

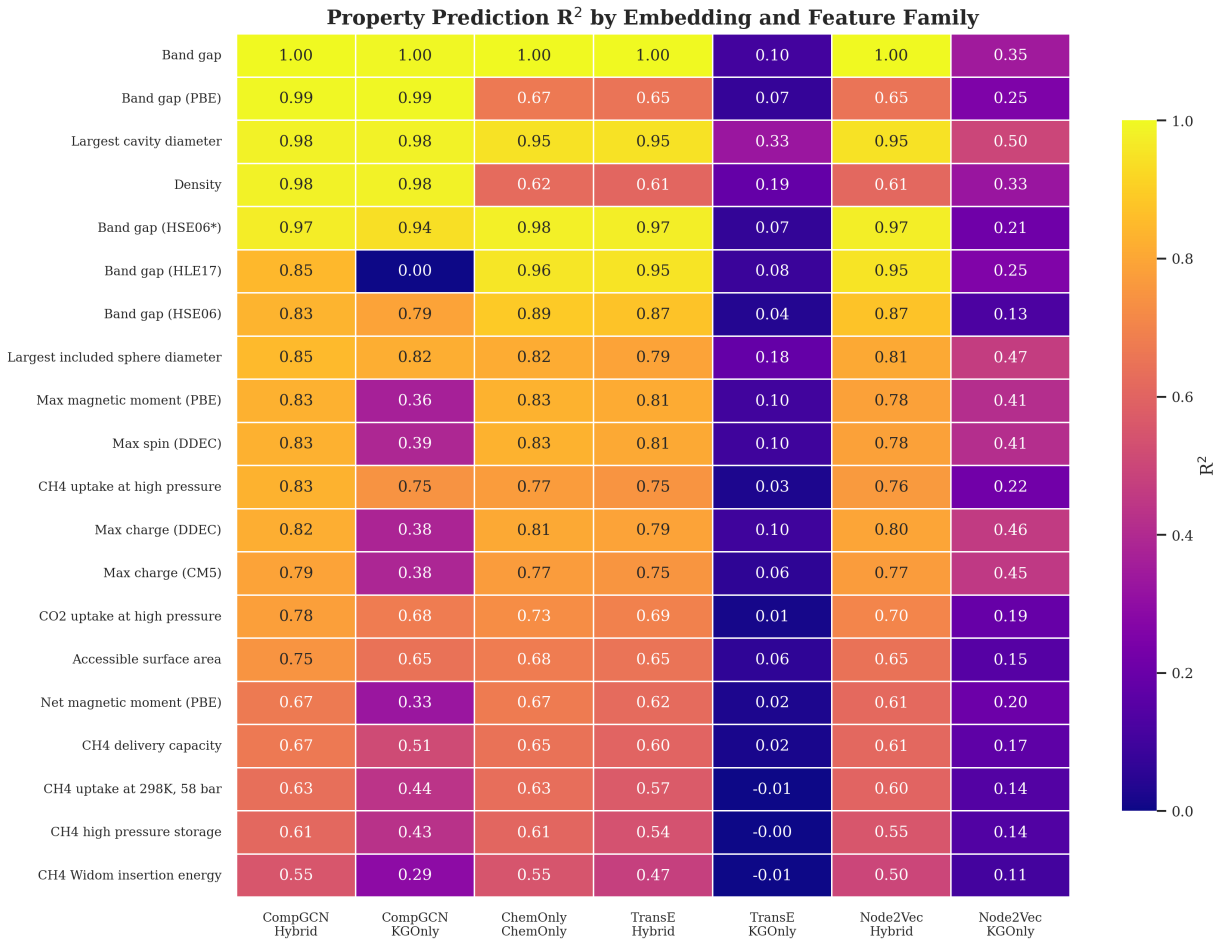

Figure 8. Property prediction R^2 heatmap. Rows: target properties (top 20 by max R^2). Columns: embedding by feature family combinations.

D. Concept Vector Analysis

Figure 9 shows pairwise correlations between learned concept vectors.

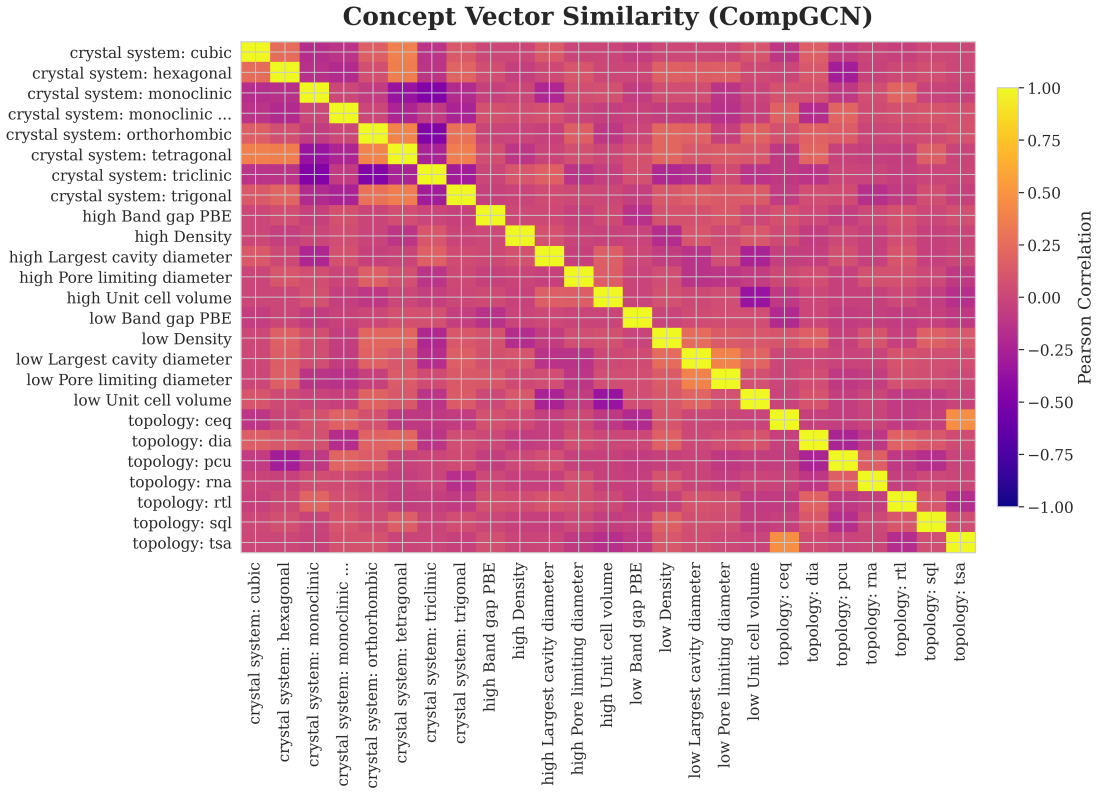

Figure 9. Concept vector similarity (Pearson correlation) for CompGCN embeddings. Crystal system, property thresholds, and metal/topology labels form distinct clusters.

E. Additional Tables

Table 8. Property reliability weights used for DAC screening. Weights in the present study were set manually by experts, and can be changed depending on the goals of the screening campaign.

Property	**Weight**
CO_2 binding energy	0.34
H_2O binding energy	0.33
Thermal stability	0.15
Density	0.10
Pore limiting diameter	0.08