22 February 2026

Stereoelectronic-aware catalyst embeddings from 2D graphs via 2D-3D multi-view alignment

Li-Cheng Xu[1], Fenglei Cao[1], Yuan Qi[2,1]

1. Shanghai Academy of Artificial Intelligence for Science

2. Artificial Intelligence Innovation and Incubation Institute Fudan University

Abstract

In catalyst discovery, data-driven design and screening are playing an increasingly central role. However, molecular descriptors that encode the decisive steric and electronic properties often lack automation and unification. We present CatEmb, a stereoelectronic-aware catalyst embedding learned through contrastive alignment of 2D and 3D molecular graph representations. Trained on the curated CatCompDB dataset, CatEmb encodes implicit 3D structural and energetic information directly from a 2D molecular graph. We demonstrate its efficacy in clustering chemically diverse ligand classes and enhancing reaction performance prediction in quantitative structure-performance relationship models. Furthermore, we develop a novel, similarity-based iterative strategy using CatEmb. For high-throughput experimental campaigns aimed at exhaustively mapping a large condition space, this strategy efficiently identifies high-performing, substrate-general catalysts, significantly outperforming random and model-based approaches. When applied to large-scale virtual libraries with limited experimental validation, it successfully prioritizes high-performance catalysts tailored for specific reactions. CatEmb provides a versatile and efficient foundation for data-driven catalyst discovery.

Keywords

molecular descriptor, deep learning, catalyst discovery, contrastive learning

Stereoelectronic-aware catalyst embeddings from 2D graphs via 2D-3D multi-view alignment

Li-Cheng Xu,[1*] Fenglei Cao,[1] Yuan Qi[1,2]

[1]Shanghai Academy of Artificial Intelligence for Science, Shanghai, 200232, China

[2]Artificial Intelligence Innovation and Incubation Institute, Fudan University, Shanghai, 201203, China

*Email: xulicheng@sais.org.cn

Abstract: In catalyst discovery, data-driven design and screening are playing an increasingly central role. However, molecular descriptors that encode the decisive steric and electronic properties often lack automation and unification. We present CatEmb, a stereoelectronic-aware catalyst embedding learned through contrastive alignment of 2D and 3D molecular graph representations. Trained on the curated CatCompDB dataset, CatEmb encodes implicit 3D structural and energetic information directly from a 2D molecular graph. We demonstrate its efficacy in clustering chemically diverse ligand classes and enhancing reaction performance prediction in quantitative structure-performance relationship models. Furthermore, we develop a novel, similarity-based iterative strategy using CatEmb. For high-throughput experimental campaigns aimed at exhaustively mapping a large condition space, this strategy efficiently identifies high-performing, substrate-general catalysts, significantly outperforming random and model-based approaches. When applied to large-scale virtual libraries with limited experimental validation, it successfully prioritizes high-performance catalysts tailored for specific reactions. CatEmb provides a versatile and efficient foundation for data-driven catalyst discovery.

Introduction

Catalysis serves as the cornerstone for enhancing synthetic efficiency and selectivity[1–3], playing an indispensable role in modern industrial sectors such as pharmaceuticals[4,5], polymers[6,7], and energy materials[8,9]. The activity and selectivity of catalytic reactions are fundamentally governed by the steric and electronic properties of catalyst molecules[10–12]. Consequently, the identification and optimization of suitable catalyst molecules remain the core objective and often the rate-limiting step in novel catalytic system development[13–15].

In recent years, data-driven high-throughput catalyst screening and quantitative structure-performance relationship (QSPR) modeling have been widely and successfully applied to develop new catalytic systems[16–18]. This strategy hinges on transforming chemical reaction data, particularly that of catalyst molecules, into digital representations[19,20]. Machine learning or deep learning models are then employed to establish quantitative relationships between the reaction system and its performance metrics (e.g., reactivity, selectivity)[21–27]. The trained models can subsequently predict the performance of untested reaction combinations, thereby assisting chemists in rapidly identifying promising candidates from vast molecular libraries[17,28].

Within this pipeline, how to encode chemical reactions, and especially the decisive catalyst and ligand molecules, constitutes the critical bridge connecting the real world with virtual predictive models. Thanks to long-standing contributions from theoretical chemistry, a variety of molecular descriptors for catalysts have been developed[16,19,20]. Steric effects are often quantified using descriptors such as Sterimol parameters for substituent bulk[29] (Fig. 1A),

SPMS for van der Waals surface shape[30] (Fig. 1B), buried volume for ligand spatial occupancy around a central atom[31], and conformation-aware descriptors like ASO[32] and the Boltzmann-weighted wSterimol[33]. On the other hand, features derived from quantum chemical calculations, such as molecular vibrational frequencies (Fig. 1C) and atomic charges, are employed to characterize electronic effects[34,35]. Originally designed primarily for mechanistic studies[36,37], these descriptors effectively abstract molecules into numerical encodings reflecting their key physicochemical properties, and thus are also widely adopted as molecular representations in data-driven modeling[38–41].

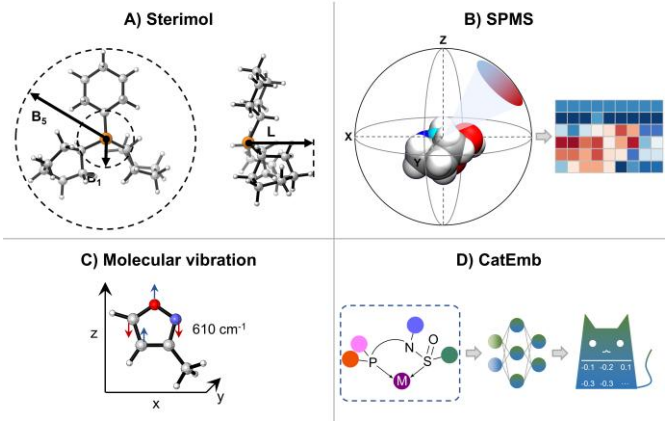

Figure 1 Representative molecular descriptors for encoding catalysts and ligands. A) Sterimol parameters for quantifying substituent steric bulk. **B)** SPMS descriptor for accurately characterizing molecular van der Waals surfaces. **C)** Example of a descriptor for electronic effects: molecular vibrational frequencies. **D)** The deep learning-based, end-to-end generated molecular descriptor CatEmb proposed in this work.

Despite their established value in data-driven modeling, the application of these descriptors faces inherent limitations. First, many steric descriptors require manual specification of atomic indices. This reduces their generalizability and impedes automation in large-scale encoding of catalysts with diverse structural scaffolds. Second, there is a lack of a unified descriptor capable of simultaneously and holistically representing both steric and electronic properties; researchers often resort to concatenating multiple distinct features, which increases model complexity and may introduce redundancy or noise.

To address these limitations and develop an end-to-end, universally applicable representation for catalysts, we drew inspiration from emergent deep-learning-based chemical embeddings such as rxnfp[42] and RXNEmb[43]. In this work, we introduce CatEmb (Fig. 1D), a 2D global catalyst descriptor informed by stereoelectronic landscapes. By employing a dual-stream architecture that aligns embeddings from 2D and 3D molecular graph neural networks, CatEmb distills 3D geometric information and single point energies into a highly efficient 2D graph embedding space. This allows the model to derive implicit steric and electronic features directly from molecular SMILES, bypassing the need for cumbersome conformational searches or time-consuming quantum chemical calculations during inference. We showcase the application of CatEmb in three tasks: assessing molecular similarity across broad catalyst and ligand libraries, building quantitative structure-performance models, and implementing a novel, similarity-based catalyst recommendation strategy. The results confirm that CatEmb effectively encapsulates critical stereoelectronic properties, highlighting its potential as a unified representation to accelerate high-throughput, data-driven catalyst discovery.

Methods

Data collection and processing

To ensure sufficient data for model training, a dedicated dataset was constructed through a two-stage pipeline. First, we integrated several open-source, curated catalyst and ligand datasets, including Kraken[44], CLC-DB[45], OSCAR[46], and the SadPhos-Library[47], and supplemented them with additional catalyst and ligand SMILES strings collected from the literature[48] and the commercial chemical repository. After canonicalization, merging, and deduplication, an initial dataset containing 12,797 unique catalyst/ligand SMILES was obtained (Fig. 2A). Next, a privileged-scaffold template-matching strategy was applied to identify ligands featuring specific privileged scaffolds from this set. These ligands were then algorithmically coordinated with commonly associated transition metals to generate SMILES data for ligand–transition-metal complexes, yielding 53,867 unique ligand–metal complex entries. The integration of data from both stages yielded a unified dataset, CatCompDB, which comprises 66,664 entries encompassing catalysts, ligands, and their derived complexes.

To acquire 3D geometric structures and their corresponding electronic energies, each SMILES was converted into a 2D molecular graph and a reasonable initial 3D conformation was generated using RDKit[49]. These preliminary conformers were then refined through constrained geometry optimization at the GNF2-xTB level with the semi-empirical quantum-chemical package xTB[50,51], yielding optimized 3D structures and their computed energies for each 2D molecular graph (Fig. 2B). The resulting triple of 2D graph, optimized 3D structure, and its energy served as inputs and labels for the subsequent contrastive model training. The full processed dataset was subsequently partitioned into training and validation subsets in a 9:1 ratio.

Representation alignment via contrastive learning

Built upon the CatCompDB dataset, we designed a self-supervised learning framework to align molecular 2D and 3D representations through contrastive learning[52,53]. The architecture, illustrated in Fig. 2C, is inspired by Yang's work on end-to-end deep learning-based reaction representations and by studies on aligning 1D and 3D molecular representations[54].

The framework employs a dual-stream encoder architecture. The 2D encoder uses our previously developed 2D molecular graph neural network[27] to encode molecular graph of catalysts and ligand–metal complexes, producing 2D molecular embeddings. The 3D encoder utilizes Equiformer[55], a higher-degree equivariant 3D graph neural network, to encode the optimized 3D molecular structures, yielding 3D molecular embeddings. We employed the regression head inherent to the Equiformer architecture, with the auxiliary objective of fitting the xTB-computed molecular energy. This ensures that the resulting 3D embeddings encode both the geometric structure and the electronic properties of the molecules.

The contrastive learning is driven by the joint optimization of two objectives. First, an EBM (Energy-Based Model) loss[56] performs instance-level contrastive learning. It explicitly distinguishes matched 2D-3D embedding pairs from mismatched ones, pulling the embeddings of the same molecule closer in the representation space while pushing apart those of different molecules. Second, distribution-level alignment is enforced by minimizing the Kullback–Leibler divergence[57] between the sets of 2D and 3D embeddings, constraining the overall statistics of the two representation spaces. This design enables the 2D graph network to learn steric and electronic features that are

typically accessible only from 3D structures, using 2D graph input alone. The final output is CatEmb, a universal, stereoelectronic-aware catalyst representation that can be generated end-to-end directly from a 2D graph, resulting in a fixed-length 32-dimensional descriptor.

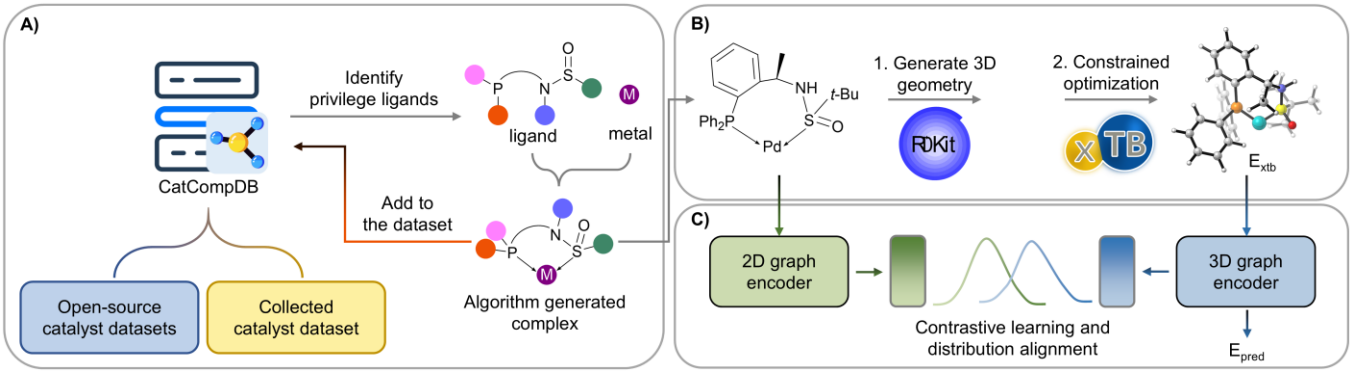

Figure 2 Generation framework of the CatEmb descriptor. A) Construction of the CatCompDB dataset via multi-source data collection and algorithmic generation of catalyst, ligand, and metal–ligand complex SMILES for model training. **B)** Conversion of SMILES from CatCompDB into 2D molecular graphs using RDKit, followed by 3D structure generation and constrained optimization with xTB to obtain optimized geometries and corresponding energies. **C)** Training of 2D and 3D molecular encoders using the 2D graphs and 3D structures/energies from CatCompDB; the trained 2D encoder is subsequently used to encode molecular graphs and produce CatEmb representations.

Results and discussion

Composition and construction of the CatCompDB dataset

The construction of the CatCompDB dataset began with the collection and integration of catalyst and ligand molecules from multiple sources. Fig. 3A shows the distribution of the collected data by source, which includes entries from publicly available datasets such as OSCAR, CLC-DB, Kraken, and SadPhos-Library, as well as manually curated entries from the literature and commercial chemical repositories. Following canonicalization using RDKit, merging, and deduplication of all collected SMILES strings, a total of 12,797 unique catalyst and ligand molecules were obtained, forming the foundational dataset for the first stage.

To generate data on ligand–metal complexes that are more relevant to catalytic chemistry, we defined ten privileged ligand types based on scaffolds commonly encountered in catalytic reactions[15,58–64]. These were identified within the foundational dataset using corresponding SMARTS pattern matching. Their quantitative distribution is shown in Fig. 3B, and representative scaffold structures are illustrated in Fig. 4A. Subsequently, an algorithm was used to coordinate these identified ligands (excluding phosphoric acid) with metals they are frequently paired with in catalysis, generating ligand–metal complexes. These structures are often considered key reactive intermediates in catalytic cycles[17]. This stage yielded 53,867 unique complexes (Fig. 3C). The integration of data from both stages resulted in a combined set of 66,664 unique molecular structures. Following geometry optimization at the GFN2-xTB level, this process yielded a final curated dataset of 62,755 optimized 3D structures along with their computed single-point energies.

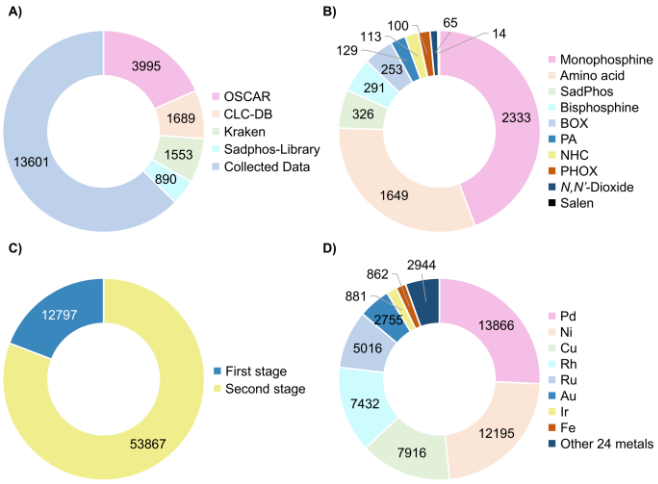

Figure 3 Data composition and statistics of the CatCompDB dataset. A) Distribution of molecular entries by data source. **B)** Abundance of the ten privileged ligand types identified in the collected data. BOX, bisoxazoline; PA, phosphoric acid; NHC, N-heterocyclic carbene; PHOX, phosphinooxazoline. **C)** Proportion of collected molecules versus algorithmically generated complexes. **D)** Distribution of central metals in the algorithmically generated ligand–metal complexes.

Owing to the broad coordination ability of the *N,N'*-dioxide ligand (reported to coordinate with 26 different metals in the literature[48]), the resulting dataset encompasses a diverse set of 32 distinct central metals (Fig. 3D). Complexes with palladium constitute the largest proportion, reflecting both the central role of palladium in numerous catalytic reactions and the general compatibility of the identified ligands with this metal. The SMARTS templates used for ligand identification and the list of their associated common coordinating metals are provided in Supplementary Table S1.

Evaluation of CatEmb for ligand similarity assessment

The trained 2D molecular graph encoder was used to generate the CatEmb representation for any molecule. We first evaluated the capability of CatEmb to assess ligand similarity. CatEmb descriptors were generated for the ten privileged ligand types identified via SMARTS patterns (Fig. 4A) and projected into a 2D plane using t-SNE[65] (Fig. 4B). Different colors and shapes denote different ligand types. Note that for t-SNE visualization, each ligand was assigned to a single class to avoid data point duplication, whereas the counts in Fig. 3B allow for multi-class membership. The results show that ligands of the same type cluster closely together, while types with distinct steric and electronic properties, such as phosphines, amino acids, and BOX ligands, occupy separate regions in the projection space.

Monophosphine ligands, the most numerous and diverse category, occupy the largest area in the 2D projection (left panel of Fig. 4C). The 1,941 monophosphine samples in the dataset encompass 40.63% of the total projection area (area calculation details provided in Supplementary Information Section 3), highlighting their extensive diversity in both steric bulk and electronic properties. This aligns with the widespread application of monophosphines in transition-metal catalysis, owing to their highly tunable structures[44,64,66].

In contrast, *N,N'*-dioxide ligands, despite a sample size of 65, occupy the smallest area in the projection, accounting for only 0.20% of the total (middle panel of Fig. 4C). This limited spatial dispersion arises because the *N,N'*-dioxide

ligands in CatCompDB were primarily derived from a study[48] focused on a specific reaction (Michael addition), resulting in a narrow range of steric and electronic variation. Their distinct coordination mode and electronic character separate them from other ligand classes, forming an isolated region in the projection space.

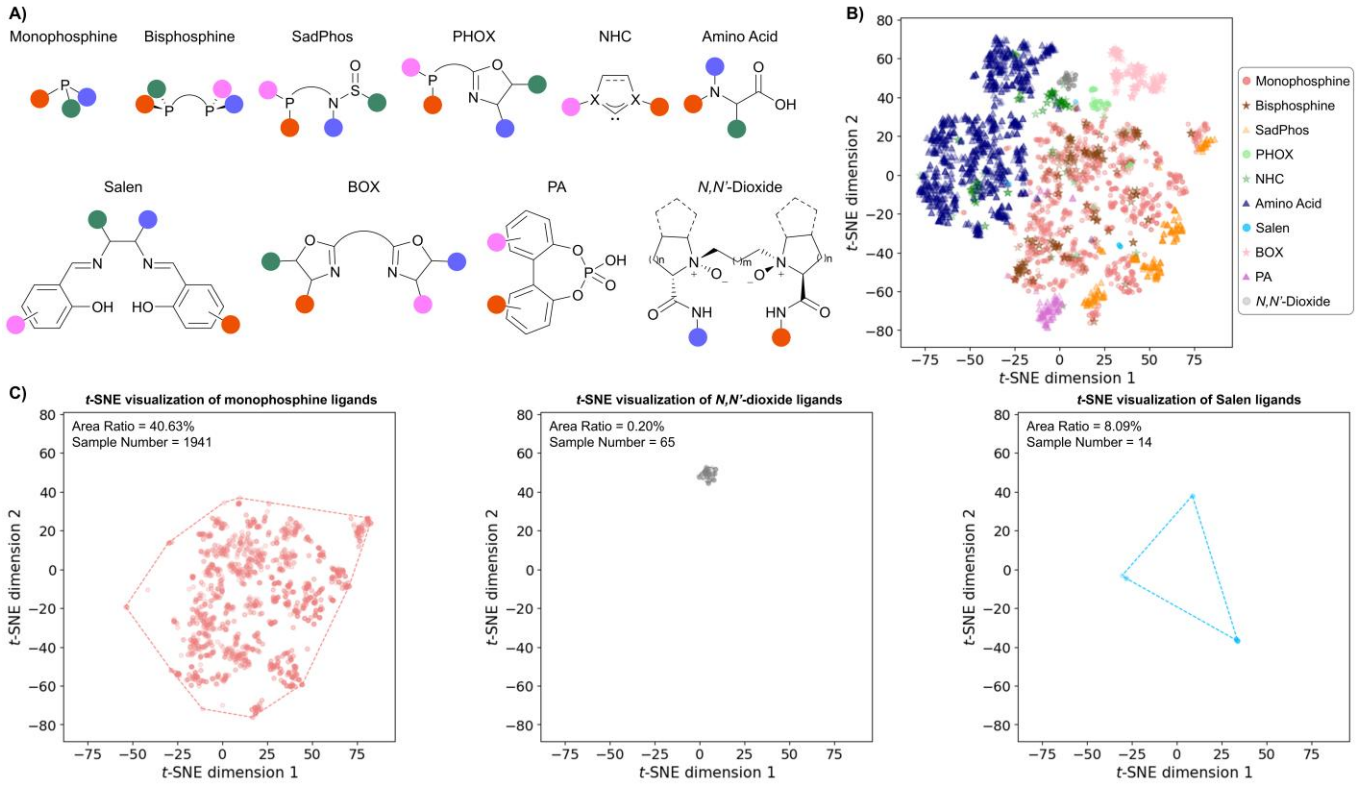

Figure 4 Analysis of ligand diversity in CatEmb space via t-SNE projection reveals steric-electronic clustering patterns. A) Major scaffolds of the ten privileged ligand classes studied. **B)** Joint t-SNE projection of CatEmb codes for all ligand classes, with each class color-coded. **C)** Individual t-SNE projections for three representative classes: Monophosphine (largest spread, highest sample count), N,N'-Dioxide (smallest spread), and Salen (smallest sample count), illustrating the correlation between chemical diversity and latent space distribution.

Salen ligands are the least represented among the ten types, with only 14 samples. However, they form 3 distinct clusters in the t-SNE plot due to significant variations introduced by their substituents. The first cluster comprises the base Salen scaffold bearing only phenolic hydroxyl groups. The second cluster incorporates additional electron-donating alkyl substituents (e.g., methyl or tert-butyl groups) on the phenyl rings. The third cluster features saturated nitrogen-containing six-membered heterocycles (e.g., piperidine or morpholine) as substituents. The distinction arises because alkyl substituents (second cluster) provide moderate electron donation primarily through inductive effect, with modest steric demand. The *N*-heterocyclic substituents (third cluster) impart both stronger electron-donating character and greater steric bulk and rigidity owing to their larger, cyclic structure. The first cluster serves as the unmodified reference with neutral electronic and steric character. These critical physicochemical differences explain their separation into three distinct regions within the CatEmb latent space, despite sharing a common scaffold. Detailed structures of the 3 Salen types and individual t-SNE plots for the remaining seven ligand classes are provided in Supplementary Information Section 3 and Supplementary Figs. S2-S12.

Collectively, these t-SNE visualizations demonstrate that CatEmb effectively encodes and discriminates subtle steric and electronic differences among ligands. It not only accurately clusters established ligand classes but also sensitively captures property gradients within a shared scaffold induced by substituent variations. This establishes it as a translator that converts accessible 2D structural information into rich stereoelectronic profiles derived from 3D molecular characteristics.

Application of CatEmb in predictive modeling and catalyst recommendation

Next, we systematically evaluated the application of CatEmb in QSPR modeling and catalyst recommendation using 3 reaction datasets rich in catalyst/ligand diversity. We selected the palladium-catalyzed C–H arylation of imidazoles reported by Doyle[67] (Fig. 5A) and the asymmetric thiol addition reported by Denmark[32] (Fig. 5B) to build quantitative models for reaction reactivity and enantioselectivity, respectively. For each dataset, ligands or catalysts were encoded using CatEmb. The performance of CatEmb was tested by pairing it separately with 3 established reaction descriptors, namely the deep learning-based rxnfp[42] and RXNEmb[43], and the rule-based DRFP[68]. For each dataset, a random split (8:2 for C–H arylation; 675:400 for thiol addition) was applied, and the entire modeling procedure was repeated 10 times to ensure robust evaluation. Modeling with an ExtraTrees regressor[69] showed that the "rxnfp + CatEmb" strategy performed best. For the C–H arylation, the average test-set R^2 and MAE for yield prediction were 0.759 and 9.24%, respectively. For the thiol addition, the average R^2 and MAE for $\Delta\Delta G$ prediction were 0.902 and 0.147 kcal/mol. Representative regression plots are shown in the lower-left panels of Fig. 5A and 5B. Notably, incorporating CatEmb consistently and significantly improved the performance of models using reaction descriptors alone (see Supplementary Information Section 4 for complete results).

Beyond QSPR modeling, we focused on CatEmb's potential for recommending "general-purpose" optimal catalysts. The identification of robust, substrate-general conditions presents a distinct challenge. This is exemplified by Doyle's prior work[67], which used bandit optimization-based reinforcement learning to identify the optimal ligand for C–H arylation—defined as the ligand with the highest average yield across all substrate combinations (Fig. 5A). While conventional model-based optimization strategies excel at tailoring conditions for limited substrates[70,71], they are often less effective at discovering catalysts that perform well across diverse substrates. Inspired by this challenge and guided by the observation that privileged ligands often share similar stereoelectronic properties, we designed an iterative recommendation strategy based directly on CatEmb similarity: starting from a small random set, each iteration selects the untested ligands most similar in CatEmb space to the best-performing ligand identified so far.

Applied to C–H arylation dataset (24 ligands × 64 substrate combinations), the CatEmb-similarity-based strategy (Fig. 5A, lower right, blue line) identified the optimal ligand Cy-BippyPhos in over 60 of the 100 independent trials after testing only 10 ligands (i.e., 640 total reactions). It significantly outperformed both random selection (orange line) and a greedy QSPR-model-based strategy (green line). Applied to the asymmetric thiol addition dataset (43 catalysts × 25 substrate combinations), the same strategy (Fig. 5B, lower right, blue line) also showed a clear advantage: the success rate for identifying the optimal catalyst increased sharply after testing only 10 catalysts across 100 trials. These results demonstrate that the steric-electronic similarity encoded by CatEmb can effectively guide the discovery of robust, substrate-general catalysts. Implementation details for the three catalyst recommendation strategies are provided in Supplementary Information, Section 5.

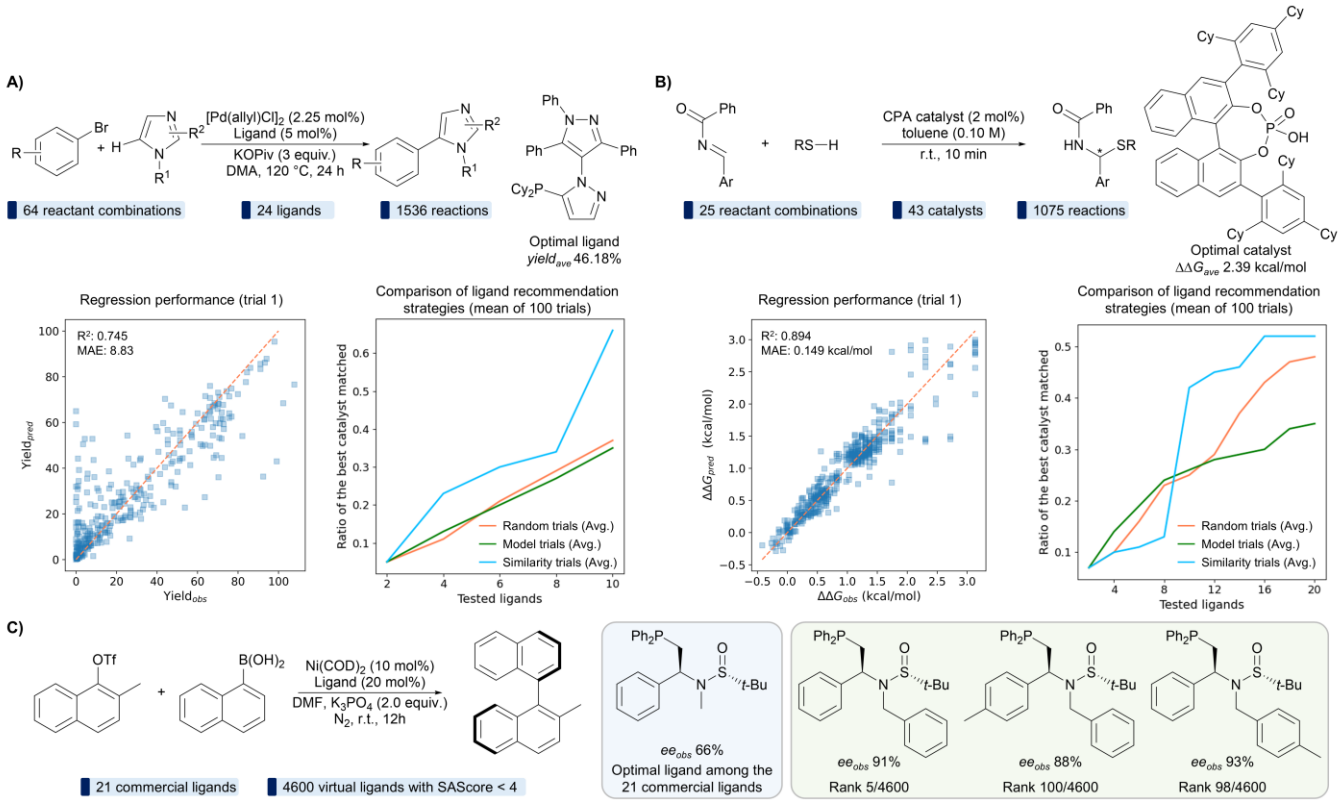

Figure 5 Application of CatEmb in reaction performance prediction and catalyst/ligand recommendation. A) Performance in yield QSPR modeling and ligand recommendation for palladium-catalyzed C–H arylation of imidazoles. **B)** Performance in enantioselectivity (ΔΔG) QSPR modeling and catalyst recommendation for asymmetric thiol addition. **C)** Ligand recommendation for Ni-catalyzed atroposelective Suzuki–Miyaura cross-coupling from a large virtual library, using a limited experimental dataset.

To further test the strategy in a scenario involving a large virtual library with limited experimental validation, which is a common step in data-driven catalyst development[72,73], we applied it to the ligand screening for the Ni-catalyzed atroposelective Suzuki–Miyaura cross-coupling reported by Hong[17]. The original study built a predictive model by applying transfer learning from literature Pd-catalysis data to 21 commercially available Ni-catalysis ligands (Fig. 5C, blue panel). This model was then used to virtually screen over 10,000 candidates, ultimately identifying 3 high-enantioselectivity ligands for experimental validation, with *ee* values of 91%, 88%, and 93%, respectively. We computed CatEmb descriptors for the Ni complexes of a subset of 4,600 synthesizable candidates from that virtual library and ranked them by CatEmb distance to the best-performing ligand among the initial 21. The 3 experimentally confirmed top performers were ranked 5th, 100th, and 98th within the 4,600 candidates (Fig. 5C, green panel). This confirms that ranking by CatEmb similarity alone can potentially prioritize high-performance ligands from vast virtual libraries for experimental testing.

Conclusion

In this work, we have developed CatEmb, a novel, stereoelectronic-aware molecular representation designed to overcome the limitations of traditional descriptors in catalyst informatics. By constructing the comprehensive CatCompDB dataset and employing a contrastive learning framework that aligns embeddings from 2D and 3D

molecular graphs, CatEmb successfully distills implicit 3D geometric and electronic information into a compact descriptor generated directly from 2D molecular graphs. This end-to-end approach bypasses the need for costly conformational searches or quantum-chemical calculations during inference, offering a unified and automated alternative to manual feature engineering or concatenation of multiple descriptors.

We have demonstrated the efficacy of CatEmb across multiple critical tasks in data-driven catalyst discovery. First, it provides a chemically intuitive similarity metric, as evidenced by its ability to differentiate and cluster diverse ligand classes based on subtle steric-electronic variations in a low-dimensional latent space. Second, when integrated as a molecular feature, CatEmb consistently enhances the predictive performance of QSPR models for key catalytic outcomes, including reaction yield and enantioselectivity. Most importantly, we designed a novel, similarity-based iterative recommendation strategy leveraging CatEmb. This strategy proves highly effective for two distinct discovery paradigms: efficiently identifying robust, substrate-general catalysts within high-throughput experimental campaigns, and successfully prioritizing high-performance, reaction-specific candidates from vast virtual libraries with limited prior experimental data. While the current validations were performed within specific ligand classes, the fundamental design of CatEmb relies on stereoelectronic features for similarity assessment. This core mechanism theoretically supports its future extension to recommendations that operate across ligand scaffold types. Collective results demonstrate that the stereoelectronic similarity captured by CatEmb provides a powerful and efficient heuristic for catalyst exploration and prioritization.

Collectively, our results establish CatEmb as a versatile and powerful foundational tool for data-driven catalyst discovery. It bridges the gap between easily accessible 2D molecular structures and the rich, decisive stereoelectronic profiles that govern catalytic performance. By providing an efficient, accurate, and generalizable representation, CatEmb has the potential to significantly accelerate the exploration and optimization of catalytic chemical space.

Author contributions

L.-C.X.: conceptualization, methodology, software, validation, investigation, data curation, funding acquisition, writing - original draft, writing - review and editing, and visualization. F.C.: supervision and resources. Y.Q.: supervision, resources, and funding acquisition.

Conflicts of interest

There are no conflicts to declare.

Data availability

The CatCompDB dataset is available via https://github.com/licheng-xu-echo/CatEmb. Codes for data preprocessing, model training, descriptor generation are available via https://github.com/licheng-xu-echo/CatEmb.

Acknowledgements

Generous support by the National Natural Science Foundation of China (82394432 and 92249302, Y.Q.), the Shanghai Municipal Science and Technology Major Project (Grant No. 2023SHZDZX02, Y.Q.); Oriental Talent Youth Project (L.-C.X.). the Shanghai AI for Science Hundred Teams Hundred Projects (F.C.). The DFT computations and model training in this research were performed using the Inspire platform at SAIS.

Reference

1. Noyori, R. Asymmetric Catalysis: Science and Opportunities (Nobel Lecture). *Angew. Chem. Int. Ed.* **41**, 2008–2022 (2002).

2. Trost, B. M. Asymmetric catalysis: An enabling science. *Proc. Natl. Acad. Sci.* **101**, 5348–5355 (2004).

3. Armor, J. N. A history of industrial catalysis. *Catal. Today* **163**, 3–9 (2011).

4. Kar, S., Sanderson, H., Roy, K., Benfenati, E. & Leszczynski, J. Green Chemistry in the Synthesis of Pharmaceuticals. *Chem. Rev.* **122**, 3637–3710 (2022).

5. Yang, H., Yu, H., Stolarzewicz, I. A. & Tang, W. Enantioselective Transformations in the Synthesis of Therapeutic Agents. *Chem. Rev.* **123**, 9397–9446 (2023).

6. Chen, M., Zhong, M. & Johnson, J. A. Light-Controlled Radical Polymerization: Mechanisms, Methods, and Applications. *Chem. Rev.* **116**, 10167–10211 (2016).

7. Corrigan, N., Shanmugam, S., Xu, J. & Boyer, C. Photocatalysis in organic and polymer synthesis. *Chem. Soc. Rev.* **45**, 6165–6212 (2016).

8. Chu, S., Cui, Y. & Liu, N. The path towards sustainable energy. *Nat. Mater.* **16**, 16–22 (2017).

9. Kitanosono, T., Masuda, K., Xu, P. & Kobayashi, S. Catalytic Organic Reactions in Water toward Sustainable Society. *Chem. Rev.* **118**, 679–746 (2018).

10. Zhang, Z., Butt, N. A. & Zhang, W. Asymmetric Hydrogenation of Nonaromatic Cyclic Substrates. *Chem. Rev.* **116**, 14769–14827 (2016).

11. Zhang, S.-Q. & Hong, X. Mechanism and Selectivity Control in Ni- and Pd-Catalyzed Cross-Couplings Involving Carbon–Oxygen Bond Activation. *Acc. Chem. Res.* **54**, 2158–2171 (2021).

12. Mondal, S. *et al.* Enantioselective Radical Reactions Using Chiral Catalysts. *Chem. Rev.* **122**, 5842–5976 (2022).

13. Romero, N. A. & Nicewicz, D. A. Organic Photoredox Catalysis. *Chem. Rev.* **116**, 10075–10166 (2016).

14. Holmberg-Douglas, N. & Nicewicz, D. A. Photoredox-Catalyzed C–H Functionalization Reactions. *Chem. Rev.* **122**, 1925–2016 (2022).

15. Li, W. & Zhang, J. Sadphos as Adaptive Ligands in Asymmetric Palladium Catalysis. *Acc. Chem. Res.* acs.accounts.3c00648 (2024) doi:10.1021/acs.accounts.3c00648.

16. Zahrt, A. F., Athavale, S. V. & Denmark, S. E. Quantitative Structure–Selectivity Relationships in Enantioselective Catalysis: Past, Present, and Future. *Chem. Rev.* **120**, 1620–1689 (2020).

17. Xu, X.-Y., Liu, L.-G., Xu, L.-C., Zhang, S.-Q. & Hong, X. Transfer Learning-Enabled Ligand Prediction for Ni-Catalyzed Atroposelective Suzuki–Miyaura Cross-Coupling Based on Mechanistic Similarity: Leveraging Pd Knowledge for Ni Discovery. *J. Am. Chem. Soc.* **147**, 15318–15328 (2025).

18. Gallarati, S., Bucci, E. M., Doyle, A. G. & Sigman, M. S. Transferable enantioselectivity models from sparse data. *Nature* https://doi.org/10.1038/s41586-026-10239-7 (2026) doi:10.1038/s41586-026-10239-7.

19. Gallegos, L. C., Luchini, G., St. John, P. C., Kim, S. & Paton, R. S. Importance of Engineered and Learned Molecular Representations in Predicting Organic Reactivity, Selectivity, and Chemical Properties. *Acc. Chem. Res.* **54**, 827–836 (2021).

20. Oliveira, J. C. A. *et al.* When machine learning meets molecular synthesis. *Trends Chem.* **4**, 863–885 (2022).

21. Moskal, M., Beker, W., Szymkuć, S. & Grzybowski, B. A. Scaffold-Directed Face Selectivity Machine-Learned from Vectors of Non-covalent Interactions. *Angew. Chem. Int. Ed.* **60**, 15230–15235 (2021).

22. Reid, J. P., Betinol, I. O. & Kuang, Y. Mechanism to model: a physical organic chemistry approach to reaction prediction. *Chem. Commun.* **59**, 10711–10721 (2023).

23. Xu, L.-C. *et al.* Enantioselectivity prediction of pallada-electrocatalysed C–H activation using transition state knowledge in machine learning. *Nat. Synth.* **2**, 321–330 (2023).

24. Zhang, S. *et al.* Bridging Chemical Knowledge and Machine Learning for Performance Prediction of Organic Synthesis. *Chem. – Eur. J.* **29**, e202202834 (2023).

25. Ruos, M. E. *et al.* Data Science-Guided Development of Deoxyfluorination Reagents with Enhanced Reactivity, Practicality, and Safety. *J. Am. Chem. Soc.* **147**, 25815–25824 (2025).

26. Dalmau, D., Sigman, M. S. & Alegre-Requena, J. V. Machine learning workflows beyond linear models in low-data regimes. *Chem. Sci.* **16**, 8555–8560 (2025).

27. Xu, L.-C., Tang, M.-J., An, J., Cao, F. & Qi, Y. A unified pre-trained deep learning framework for cross-task reaction performance prediction and synthesis planning. *Nat. Mach. Intell.* **7**, 1561–1571 (2025).

28. Ruos, M. E. *et al.* Data Science-Guided Development of Deoxyfluorination Reagents with Enhanced Reactivity, Practicality, and Safety. *J. Am. Chem. Soc.* **147**, 25815–25824 (2025).

29. Verloop, A., Hoogenstraaten, W. & Tipker, J. Development and Application of New Steric Substituent Parameters in Drug Design. in *Drug Design* 165–207 (Elsevier, 1976). doi:10.1016/B978-0-12-060307-7.50010-9.

30. Xu, L.-C. *et al.* A Molecular Stereostructure Descriptor Based On Spherical Projection. *Synlett* **32**, 1837–1842 (2021).

31. Hillier, A. C. *et al.* A Combined Experimental and Theoretical Study Examining the Binding of *N* -Heterocyclic Carbenes (NHC) to the Cp*RuCl (Cp* = η^5 -C_5Me_5) Moiety: Insight into Stereoelectronic Differences between Unsaturated and Saturated NHC Ligands. *Organometallics* **22**, 4322–4326 (2003).

32. Zahrt, A. F. *et al.* Prediction of higher-selectivity catalysts by computer-driven workflow and machine learning. *Science* **363**, eaau5631 (2019).

33. Brethomé, A. V., Fletcher, S. P. & Paton, R. S. Conformational Effects on Physical-Organic Descriptors: The Case of Sterimol Steric Parameters. *ACS Catal.* **9**, 2313–2323 (2019).

34. Ahneman, D. T., Estrada, J. G., Lin, S., Dreher, S. D. & Doyle, A. G. Predicting reaction performance in C–N cross-coupling using machine learning. *Science* **360**, 186–190 (2018).

35. Li, X., Zhang, S., Xu, L. & Hong, X. Predicting Regioselectivity in Radical C−H Functionalization of Heterocycles through Machine Learning. *Angew. Chem. Int. Ed.* **59**, 13253–13259 (2020).

36. Harper, K. C., Bess, E. N. & Sigman, M. S. Multidimensional steric parameters in the analysis of asymmetric catalytic reactions. *Nat. Chem.* **4**, 366–374 (2012).

37. Brethomé, A. V., Fletcher, S. P. & Paton, R. S. Conformational Effects on Physical-Organic Descriptors: The Case of Sterimol Steric Parameters. *ACS Catal.* **9**, 2313–2323 (2019).

38. Tâmega, G. S., Costa, M. O., De Araujo Pereira, A. & Barbosa Ferreira, M. A. Data Science Guiding Analysis of Organic Reaction Mechanism and Prediction. *Chem. Rec.* **24**, e202400148 (2024).

39. Escayola, S., Bahri-Laleh, N. & Poater, A. % V_{Bur} index and steric maps: from predictive catalysis to machine learning. *Chem. Soc. Rev.* **53**, 853–882 (2024).

40. Stenfors, B. A. *et al.* Conformation Dependent Features of Bisphosphine Ligands. *J. Org. Chem.* **90**, 13874–13884 (2025).

41. Cadge, J. A., Hart, S. D., Walroth, R. C., Mack, K. A. & Sigman, M. S. Bisphosphine ligand conformer selection to enhance descriptor database representation: improving statistical modelling outcomes. *Chem. Sci.* **16**, 20473–20485 (2025).

42. Schwaller, P. *et al.* Mapping the space of chemical reactions using attention-based neural networks. *Nat. Mach. Intell.* **3**, 144–152 (2021).

43. Liu, W., Cao, F., Qi, Y. & Xu, L.-C. A Pre-trained Reaction Embedding Descriptor Capturing Bond Transformation Patterns. Preprint at https://doi.org/10.48550/arXiv.2601.03689 (2026).

44. Gensch, T. *et al.* A Comprehensive Discovery Platform for Organophosphorus Ligands for Catalysis. *J. Am. Chem. Soc.* **144**, 1205–1217 (2022).

45. Yu, G. *et al.* Clc-db: an open-source online database of chiral ligands and catalysts. *J. Cheminformatics* **17**, 45 (2025).

46. Gallarati, S. *et al.* OSCAR: an extensive repository of chemically and functionally diverse organocatalysts. *Chem. Sci.* **13**, 13782–13794 (2022).

47. Yu, S. SadPhos Library: A Comprehensive Resource for Exploring Chiral Ligand Chemical Space. *Chem. – Asian J.* **20**, e202500023 (2025).

48. Tang, M. *et al.* Data-Driven Modeling of *N,N′*-Dioxide/Metal-Catalyzed Asymmetric Michael Additions. *Angew. Chem. Int. Ed.* e18560 (2025) doi:10.1002/anie.202518560.

49. RDKit: open-source chemoinformatics and machine learning. http://www.rdkit.org.

50. Bannwarth, C., Ehlert, S. & Grimme, S. GFN2-xTB—An Accurate and Broadly Parametrized Self-Consistent Tight-Binding Quantum Chemical Method with Multipole Electrostatics and Density-Dependent Dispersion Contributions. *J. Chem. Theory Comput.* **15**, 1652–1671 (2019).

51. Bannwarth, C. *et al.* Extended TIGHT-BINDING quantum chemistry methods. *WIREs Comput. Mol. Sci.* **11**, e1493 (2021).

52. Hadsell, R., Chopra, S. & LeCun, Y. Dimensionality Reduction by Learning an Invariant Mapping. in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2 (CVPR'06)* vol. 2 1735–1742 (IEEE, New York, NY, USA, 2006).

53. Oord, A. van den, Li, Y. & Vinyals, O. Representation Learning with Contrastive Predictive Coding. Preprint at https://doi.org/10.48550/arXiv.1807.03748 (2019).

54. Shi, R., Yu, G., Huo, X. & Yang, Y. Prediction of chemical reaction yields with large-scale multi-view pre-training. *J. Cheminformatics* **16**, 22 (2024).

55. Liao, Y.-L., Wood, B. M., Das, A. & Smidt, T. E. EquiformerV2: Improved Equivariant Transformer for Scaling to Higher-Degree Representations. in *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024* (OpenReview.net, 2024).

56. LeCun, Y. *et al.* A tutorial on energy-based learning. *Predict. Struct. Data* **1**, (2006).

57. Kullback, S. & Leibler, R. A. On Information and Sufficiency. *Ann. Math. Stat.* **22**, 79–86 (1951).

58. Tokunaga, M., Larrow, J. F., Kakiuchi, F. & Jacobsen, E. N. Asymmetric Catalysis with Water: Efficient Kinetic Resolution of Terminal Epoxides by Means of Catalytic Hydrolysis. *Science* **277**, 936–938 (1997).

59. Glos, M. & Reiser, O. Aza-bis(oxazolines): New Chiral Ligands for Asymmetric Catalysis. *Org. Lett.* **2**, 2045–2048 (2000).

60. Liao, G., Zhang, T., Lin, Z. & Shi, B. Transition Metal-Catalyzed Enantioselective C−H Functionalization via Chiral Transient Directing Group Strategies. *Angew. Chem. Int. Ed.* **59**, 19773–19786 (2020).

61. Connon, R., Roche, B., Rokade, B. V. & Guiry, P. J. Further Developments and Applications of Oxazoline-Containing Ligands in Asymmetric Catalysis. *Chem. Rev.* **121**, 6373–6521 (2021).

62. Shimizu, H., Nagasaki, I. & Saito, T. Recent advances in biaryl-type bisphosphine ligands. *Tetrahedron* **61**, 5405–5432 (2005).

63. Zhao, P.-F. *et al.* Recent advances in chiral phosphoric acids for asymmetric organocatalysis: a catalyst design perspective. *Org. Biomol. Chem.* **23**, 7872–7913 (2025).

64. Tolman, C. A. Steric effects of phosphorus ligands in organometallic chemistry and homogeneous catalysis. *Chem. Rev.* **77**, 313–348 (1977).

65. Maaten, L. van der & Hinton, G. Visualizing Data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).

66. Tolman, C. A. Electron donor-acceptor properties of phosphorus ligands. Substituent additivity. *J. Am. Chem. Soc.* **92**, 2953–2956 (1970).

67. Wang, J. Y. *et al.* Identifying general reaction conditions by bandit optimization. *Nature* **626**, 1025–1033 (2024).

68. Probst, D., Schwaller, P. & Reymond, J.-L. Reaction classification and yield prediction using the differential reaction fingerprint DRFP. *Digit. Discov.* **1**, 91–97 (2022).

69. Geurts, P., Ernst, D. & Wehenkel, L. Extremely randomized trees. *Mach. Learn.* **63**, 3–42 (2006).

70. Shields, B. J. *et al.* Bayesian reaction optimization as a tool for chemical synthesis. *Nature* **590**, 89–96 (2021).

71. Xu, L. *et al.* Towards Data-Driven Design of Asymmetric Hydrogenation of Olefins: Database and Hierarchical Learning. *Angew. Chem. Int. Ed.* **60**, 22804–22811 (2021).

72. Graff, D. E., Shakhnovich, E. I. & Coley, C. W. Accelerating high-throughput virtual screening through molecular pool-based active learning. *Chem. Sci.* **12**, 7866–7881 (2021).

73. Ma, S. *et al.* Data-driven discovery of active phosphine ligand space for cross-coupling reactions. *Chem. Sci.* **15**, 13359–13368 (2024).