Digital Discovery

ROYAL SOCIETY OF CHEMISTRY

PAPER

Published on 02 July 2026

Cite this: DOI: 10.1039/d6dd00089d

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

Li-Cheng Xu, *[a] Fenglei Cao[a] and Yuan Qi[ab]

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 lack unification and are often manually defined. 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 (QSPR) models. Furthermore, we develop a novel similarity-based reaction recommendation strategy using CatEmb, which serves as a robust complement to QSPR model-based approaches. For high-throughput experimental campaigns aimed at exhaustively mapping a large condition space, this strategy efficiently identifies high-performing combinations and substrate-general optimal catalysts. When applied to large-scale virtual libraries with limited experimental validation, it successfully prioritizes high-performance catalysts tailored for specific reactions, and, in a separate setting, it further enables recommendation across different ligand scaffold types. CatEmb provides a versatile and efficient tool for data-driven catalyst discovery.

Received 22nd February 2026
Accepted 2nd July 2026

DOI: 10.1039/d6dd00089d

rsc.li/digitaldiscovery

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 those of catalyst molecules, into digital representations.[19–21] Machine learning or deep learning models are then employed to establish quantitative relationships between the reaction system and its performance metrics (*e.g.*, reactivity and selectivity).[22–29] The trained models can subsequently predict the performance of untested reaction combinations, thereby assisting chemists in rapidly identifying promising candidates from vast molecular libraries.[17,26]

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[30] (Fig. 1A), SPMS for van der Waals surface shape[31] (Fig. 1B), buried volume for ligand spatial occupancy around a central atom,[32] and conformation-aware descriptors such as ASO[33] and the Boltzmann-weighted wSterimol.[34] 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.[35,36] Originally designed primarily for mechanistic studies,[34,37] 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]

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

[a]*Shanghai Academy of Artificial Intelligence for Science, Shanghai, 200232, China. E-mail: xulicheng@sais.org.cn*

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

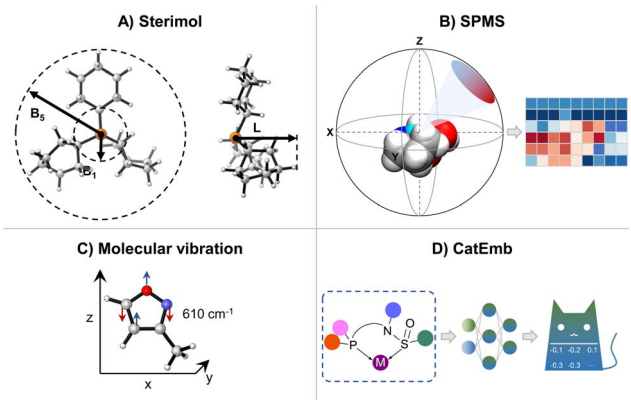

Fig. 1 Representative molecular descriptors for encoding catalysts and ligands. (A) Sterimol parameters for quantifying substituent steric bulk. (B) SPMS descriptor for accurately characterizing molecular van der Waals surfaces. (C) Example of a descriptor for electronic effects: molecular vibrational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

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 scenarios: 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 a 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 entire optimization process consumed approximately 1250 CPU core-hours. The resulting triple consisting of a 2D graph, an 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[28] to encode molecular graphs 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 framework 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, thereby constraining the overall statistics of the two representation

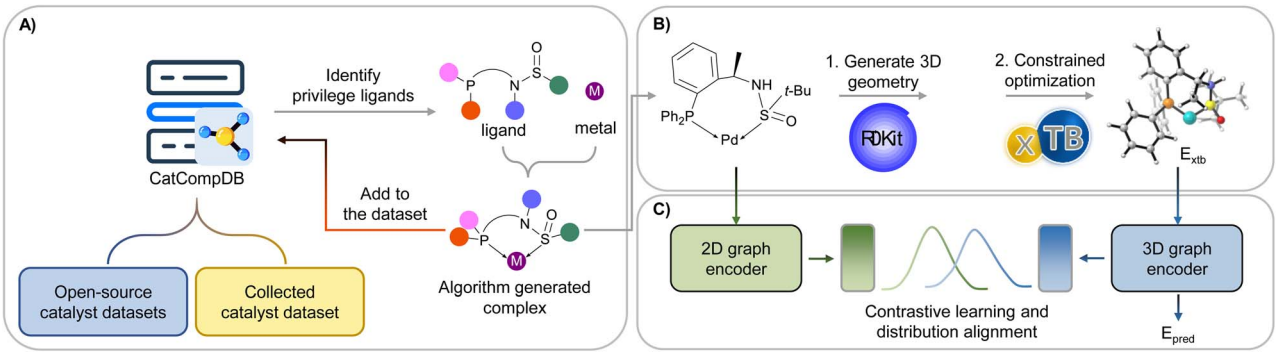

Fig. 2 Generation framework of the CatEmb descriptor. (A) Construction of the CatCompDB dataset *via* multi-source data collection and algorithmic generation of catalyst, ligand, and metal–ligand complex SMILES for model training. (B) Conversion of SMILES from CatCompDB into 2D molecular graphs using 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.

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. Training one CatEmb model with 100 epochs requires approximately 60 hours on a single NVIDIA 4090 GPU. To accommodate different task requirements, we trained CatEmb variants with dimensionalities ranging from 8 to 4096.

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.

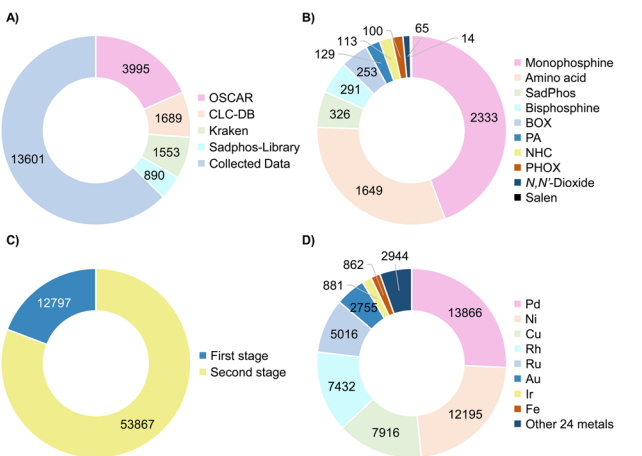

Fig. 3 Data composition and statistics of the CatCompDB dataset. (A) Distribution of molecular entries by data source. (B) Abundance of the ten privileged ligand types identified in the collected data. BOX, bisoxazoline; PA, phosphoric acid; NHC, N-heterocyclic carbene; PHOX, phosphinooxazoline. (C)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

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] The identification of these privileged ligands was guided by the classifications in the four curated catalyst and ligand datasets and their associated publications, as well as by relevant review articles on transition-metal catalysis.[15,58–60,62,63] In addition, we analysed the unclassified ligands collected from commercial chemical repositories and the literature, identifying frequently occurring scaffolds as additional privileged types. This process ultimately yielded the ten privileged ligand types. 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, which is commonly used as an organocatalyst and does not require a transition metal[63]) 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.

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

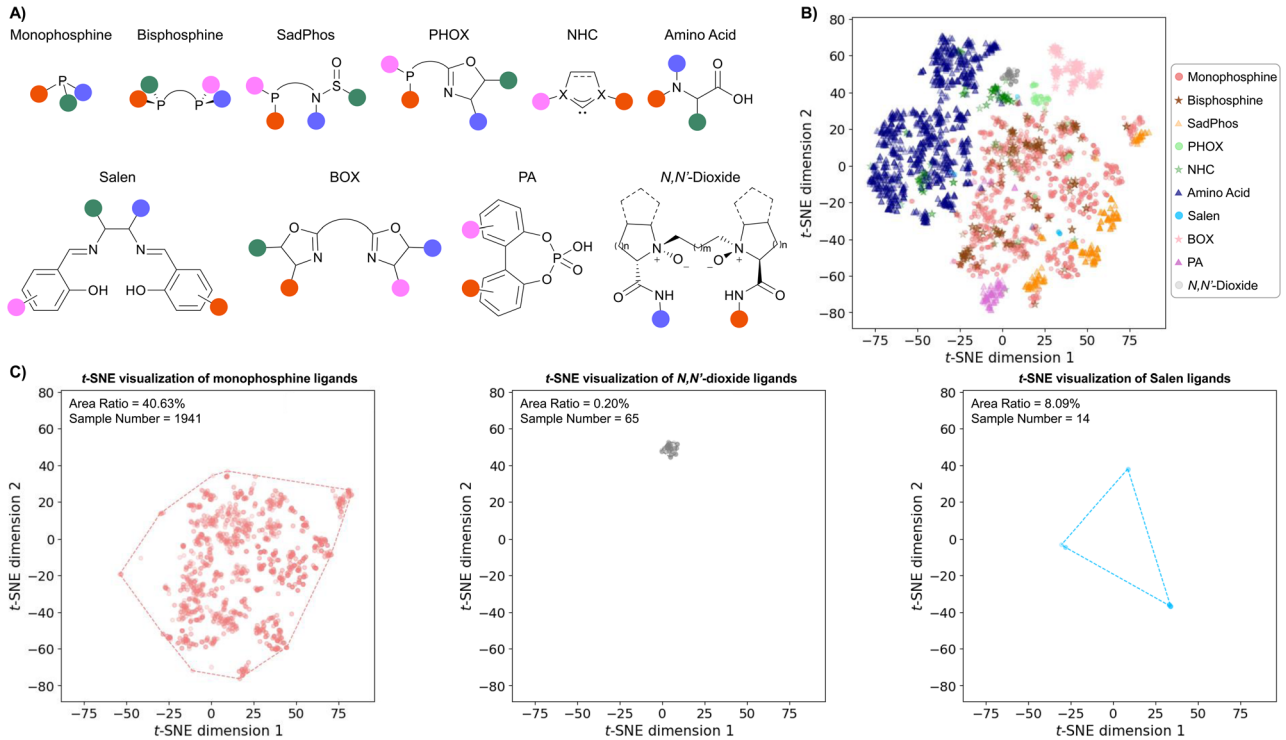

Fig. 4 Analysis of ligand diversity in CatEmb space *via* t-SNE projection reveals steric-electronic clustering patterns. (A) Major scaffolds of the ten privileged ligand classes studied. (B) Joint t-SNE projection of CatEmb codes for all ligand classes, with each class color-coded. (C) Individual t-SNE 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.

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 SI 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. After comparing the 2D–3D alignment across dimensions and considering both compactness and efficiency, we chose the 32-dimensional CatEmb for similarity evaluation (see Section 3 of the SI for details). 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 1941 monophosphine samples in the dataset encompass 40.63% of the total projection area (area calculation details provided in SI Section 4), 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.

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

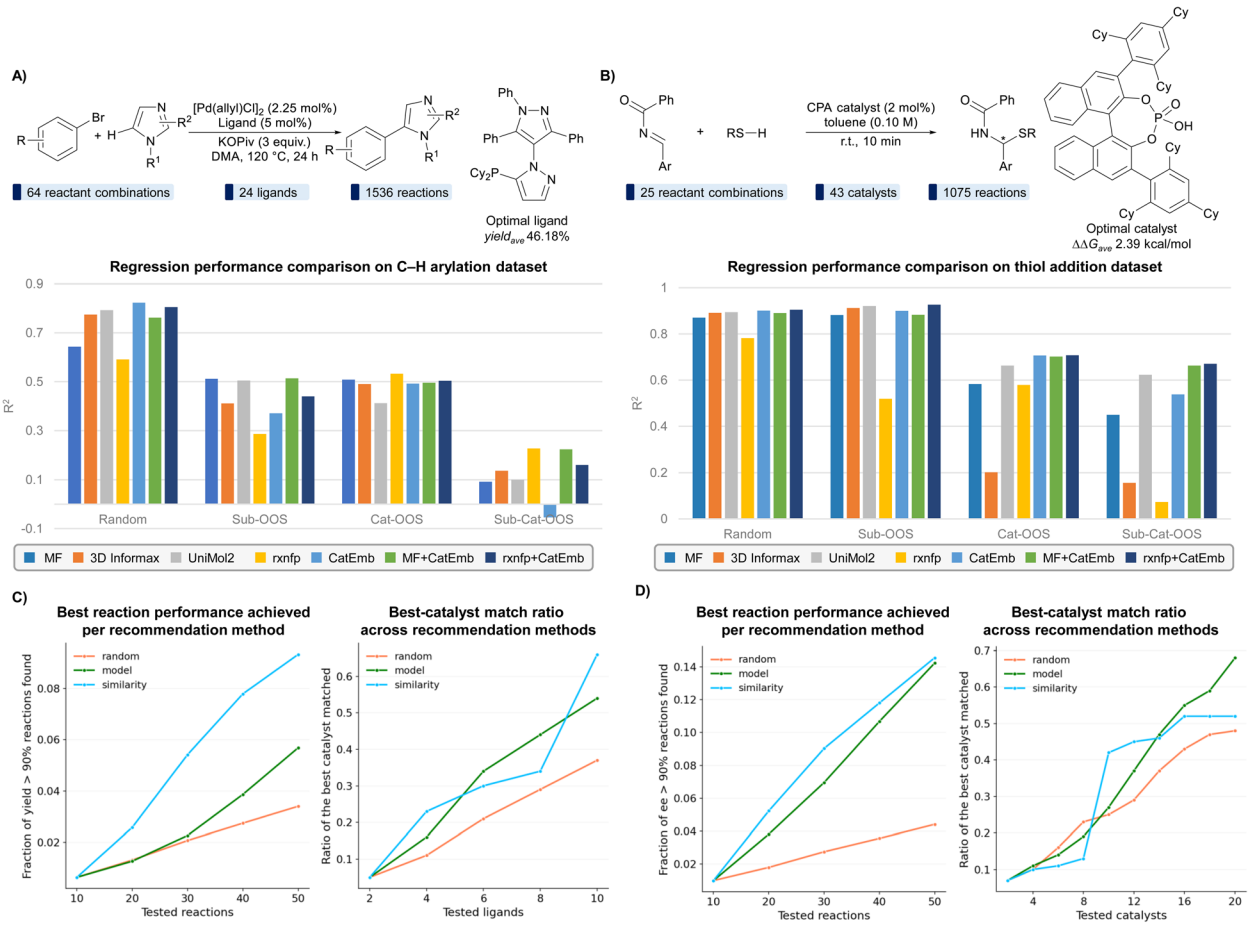

Fig. 5 Evaluation of encoding strategies for QSPR modeling and recommendation tasks on the C–H arylation and asymmetric thiol addition datasets. (A and B) QSPR performance under different splits; (C and D) comparison of random, model-based, and similarity-based recommendation strategies for high-performance conditions and optimal catalysts.

heterocycles (*e.g.*, piperidine or morpholine) as substituents. The distinction arises because alkyl substituents (second cluster) provide moderate electron donation primarily through the 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 SI Section 4 and Fig. S4–S14.

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 4 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[33] (Fig. 5B) to build quantitative models for reaction reactivity and enantioselectivity, respectively. For each dataset, we evaluated several encoding strategies under both random and out-of-sample (OOS) splits. The random split was 8 : 2 for the C–H arylation dataset and 675 : 400 for the thiol addition dataset. The OOS splits (substrate-OOS, catalyst-OOS, and substrate-catalyst combination OOS) followed the scheme originally designed for the thiol addition dataset[33] and were applied analogously to the C–H arylation dataset.

We compared three CatEmb-based reaction encoding strategies: CatEmb alone, CatEmb combined with the deep learning-based reaction fingerprint rxnfp,[42] and CatEmb combined with the Morgan fingerprint (MF),[68] where CatEmb serves as a dedicated encoder for catalysts to improve the overall reaction representation. We took MF alone as a baseline, and

additionally compared against other deep learning-based molecular descriptors (3D Infomax[69] and Uni-Mol2 (ref. 70)) as well as rxnfp alone. For each descriptor, we screened four machine learning algorithms (ExtraTrees,[71] SVR,[72] Ridge,[73] and KernelRidge[73]). For Uni-Mol2 and CatEmb, which have multiple size or dimensionality variants, we performed variant screening to optimize performance.

As shown in Fig. 5A and B, CatEmb alone achieved reasonable predictive performance, though it was not the best across all tasks. Importantly, incorporating CatEmb consistently improved prediction accuracy in most tasks, as evidenced by comparisons of MF *vs.* MF + CatEmb and rxnfp *vs.* rxnfp + CatEmb. Among all strategies, rxnfp + CatEmb exhibited the best overall performance. For the C–H arylation dataset, the R^2 values of rxnfp + CatEmb under random, sub-OOS, cat-OOS, and sub-cat-OOS splits were 0.805, 0.461, 0.498, and 0.159, respectively. For the thiol addition dataset, the corresponding R^2 values were 0.905, 0.927, 0.708, and 0.671. Detailed results are presented in SI Section 5.

Beyond QSPR modeling, we explored CatEmb's potential for catalyst and reaction condition recommendation on the C–H arylation and thiol addition datasets, focusing on two tasks: identifying high-yield or high-enantioselectivity reaction combinations and discovering robust, substrate-general optimal catalysts.

We first evaluated the ability of different strategies to efficiently identify high-performance reaction conditions. For the C–H arylation dataset (Fig. 5C, left panel), we defined high-yield reactions as those with yield >90%; for the asymmetric thiol addition dataset (Fig. 5D, left panel), high-selectivity reactions were defined as those with ee > 90%. The *x*-axis represents the number of tested reactions, and the *y*-axis represents the fraction of all high-performance reactions that have been found. We compared three recommendation strategies: random selection (orange line, baseline), a model-based strategy (green line) using a QSPR model built with the rxnfp + CatEmb descriptor combination to predict and prioritize high-performance reaction combinations, and a similarity-based strategy (blue line). Inspired by the observation that high-performance reaction conditions often share similar stereoelectronic features, this strategy recommends reaction conditions most similar in the CatEmb descriptor space to the best-performing condition identified so far. All curves are averages over 100 independent trials.

As shown in Fig. 5C (left), on the C–H arylation dataset, the model-based strategy performed poorly because the sub-cat-OOS split represents an intrinsically difficult task for QSPR on this dataset (Fig. 5A), with all tested methods yielding limited predictive accuracy. In contrast, the similarity-based strategy significantly outperformed both random selection and the model-based greedy approach. On the thiol addition dataset, the QSPR models exhibited reasonably good predictive performance under the sub-cat-OOS split (Fig. 5B), and accordingly, the model-based strategy achieved strong results (Fig. 5D, left). The similarity-based strategy performed similarly well, showing a slight advantage over the model-based approach.

We next focused on the more challenging task of identifying robust, substrate-general optimal catalysts, as previously studied by Doyle for C–H arylation[67] (optimal ligand defined as the one with the highest average yield across all substrate combinations) and Denmark for asymmetric thiol addition (optimal catalyst defined analogously). For this task, we compared the same three strategies on the right panels of Fig. 5C and D. The *x*-axis represents the number of tested ligands (or catalysts), and the *y*-axis represents the fraction of independent trials in which the current recommended set contains the optimal catalyst. Starting from a small random initial set, the similarity-based strategy (blue line) recommends untested ligands or catalysts most similar in the CatEmb space to the best-performing one identified so far.

Applied to the C–H arylation dataset (24 ligands × 64 substrate combinations, Fig. 5C right), the similarity-based strategy identified the optimal ligand Cy-BippyPhos in over 60 of the 100 trials after testing 10 ligands. The model-based greedy strategy also performed reasonably well, achieving a success rate above 50% under the same condition, which is consistent with its moderate QSPR performance on the cat-OOS split ($R^2 = 0.498$, Fig. 5A). Both strategies significantly outperformed random selection. On the asymmetric thiol addition dataset (43 catalysts × 25 substrate combinations, Fig. 5D right), where the QSPR model exhibited strong predictive performance on the catalyst-OOS split ($R^2 = 0.708$, Fig. 5B), the model-based strategy achieved a success rate of nearly 70% after testing 20 catalysts, substantially outperforming the similarity-based strategy (approximately 50% success rate). These results indicate that the model-based strategy becomes increasingly effective as QSPR performance improves, while the similarity-based strategy offers a robust alternative, particularly when predictive models suffer from poor OOS generalization. Implementation details for the three recommendation strategies are provided in SI, Section 6.

To further test the similarity-based strategy in a scenario involving a large virtual library with limited experimental validation, which is a common step in data-driven catalyst development,[74,75] 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. 6A, 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 4600 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 4600 candidates (Fig. 6A, green panel). These results suggest that ranking by CatEmb similarity may offer a useful heuristic for prioritizing candidates from large virtual libraries, though its generalizability remains to be tested.

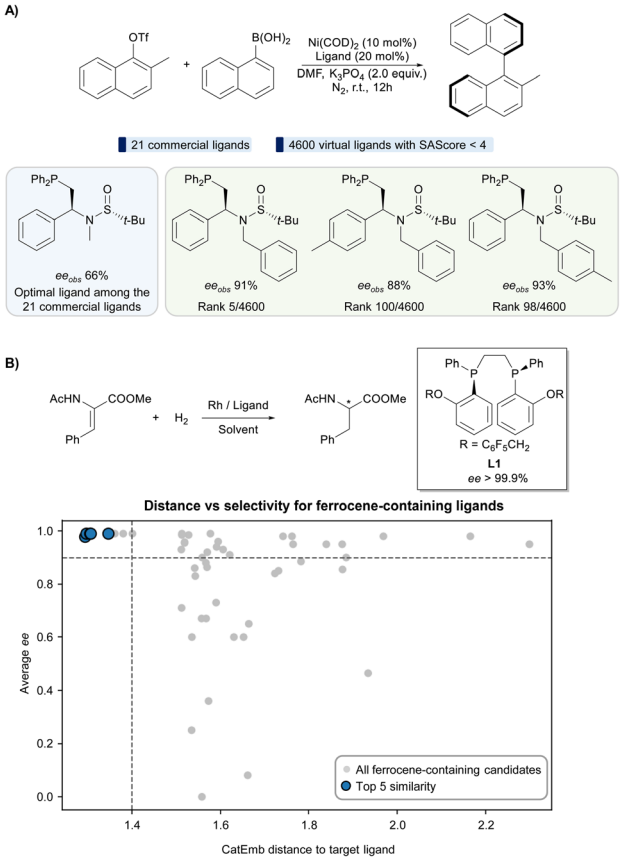

Fig. 6 CatEmb similarity for virtual library prioritization and cross-scaffold catalyst recommendation. (A) Ranking of experimentally validated ligands in a large virtual library for Ni-catalyzed Suzuki–Miyaura cross-coupling. (B) Relationship between CatEmb distance and enantioselectivity for ferrocene-containing ligands relative to a ferrocene-free target ligand.

To further test the cross-scaffold generalizability of CatEmb, we assessed whether it can facilitate catalyst recommendation across different ligand scaffolds. To this end, we utilized an asymmetric hydrogenation of olefins dataset published by Hong and colleagues,[76] which contains a diverse set of ligand types (Fig. 6B). We selected the most frequently occurring substrate, methyl (*Z*)-2-acetamido-3-phenylacrylate, and restricted the central metal to rhodium to derive a subset. The entry with the highest enantioselectivity (ee > 99.9%) employed a ferrocene-free bisphosphine ligand (**L1**). To assess cross-scaffold similarity, we computed the Euclidean distances between **L1** and all ferrocene-containing ligands in this subset. As shown in the bottom panel of Fig. 6B, the five ligands most similar to **L1** (highlighted in blue) all exhibited average ee values above 90%, while ligands with lower average selectivities showed lower similarity to **L1**. These results suggest that CatEmb similarity may offer a useful heuristic for cross-scaffold catalyst recommendation, although we note that the majority of reactions in this dataset are highly enantioselective, which makes the recommendation task inherently easier. Therefore, we suggest cautiously that CatEmb could help recommend catalysts across different ligand scaffolds, though more testing on harder tasks is needed.

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. One key aspect is that 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. Additionally, when integrated as a molecular feature, CatEmb consistently enhances the predictive performance of QSPR models for key catalytic outcomes, including reaction yield and enantioselectivity. Furthermore, we introduced a similarity-based iterative recommendation strategy leveraging CatEmb, which serves as a robust complement to QSPR model-based approaches, as validated on both high-performance reaction condition selection and optimal catalyst identification tasks. This strategy proves highly effective for two additional discovery paradigms: (1) prioritizing high-performance, reaction-specific candidates from a large virtual ligand library, where three experimentally validated ligands for Ni-catalyzed Suzuki–Miyaura cross-coupling were ranked 5th, 98th, and 100th among 4600 candidates, and (2) recommending catalysts across different ligand scaffold types, as shown by the five ferrocene-containing ligands most similar to a ferrocene-free optimal ligand, all achieving average enantioselectivities above 90% in asymmetric alkene hydrogenation. 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. Nevertheless, we note that the optimal dimensionality of CatEmb is task-dependent; while the 2D–3D alignment quality is largely insensitive to dimensionality changes (see SI Table S2 for details), QSPR performance varies considerably across dimensions (detailed in SI Fig. S3). Therefore, for new applications, users may need to perform a dimensionality search to identify the most suitable variant. Despite this current limitation, the broad availability of pre-trained CatEmb models across multiple dimensions (from 8 to 4096) facilitates such customization. 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 and trained model weights are available *via* https://doi.org/10.6084/m9.figshare.31375579. Codes for data preprocessing, model training, descriptor generation are available *via* the GitHub repository (https://github.com/licheng-xu-echo/CatEmb) and its archived Zenodo release (https://doi.org/10.5281/zenodo.20805328).

Supplementary information (SI) is available. See DOI: https://doi.org/10.1039/d6dd00089d.

Acknowledgements

Generous support from 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.), the Oriental Talent Youth Project (L.-C. X.), and the Shanghai AI for Science Hundred Teams Hundred Projects (F. C.) is gratefully acknowledged. The xTB computations and model training in this research were performed using the Inspire platform at SAIS.

Notes and references

1 R. Noyori, *Angew. Chem., Int. Ed.*, 2002, **41**, 2008–2022.
2 B. M. Trost, *Proc. Natl. Acad. Sci. U. S. A*, 2004, **101**, 5348–5355.
3 J. N. Armor, *Catal. Today*, 2011, **163**, 3–9.
4 S. Kar, H. Sanderson, K. Roy, E. Benfenati and J. Leszczynski, *Chem. Rev.*, 2022, **122**, 3637–3710.
5 H. Yang, H. Yu, I. A. Stolarzewicz and W. Tang, *Chem. Rev.*, 2023, **123**, 9397–9446.
6 M. Chen, M. Zhong and J. A. Johnson, *Chem. Rev.*, 2016, **116**, 10167–10211.
7 N. Corrigan, S. Shanmugam, J. Xu and C. Boyer, *Chem. Soc. Rev.*, 2016, **45**, 6165–6212.
8 S. Chu, Y. Cui and N. Liu, *Nat. Mater.*, 2017, **16**, 16–22.
9 T. Kitanosono, K. Masuda, P. Xu and S. Kobayashi, *Chem. Rev.*, 2018, **118**, 679–746.
10 Z. Zhang, N. A. Butt and W. Zhang, *Chem. Rev.*, 2016, **116**, 14769–14827.
11 S.-Q. Zhang and X. Hong, *Acc. Chem. Res.*, 2021, **54**, 2158–2171.
12 S. Mondal, F. Dumur, D. Gigmes, M. P. Sibi, M. P. Bertrand and M. Nechab, *Chem. Rev.*, 2022, **122**, 5842–5976.
13 N. A. Romero and D. A. Nicewicz, *Chem. Rev.*, 2016, **116**, 10075–10166.
14 N. Holmberg-Douglas and D. A. Nicewicz, *Chem. Rev.*, 2022, **122**, 1925–2016.
15 W. Li and J. Zhang, *Acc. Chem. Res.*, 2024, **57**, 489–513.
16 A. F. Zahrt, S. V. Athavale and S. E. Denmark, *Chem. Rev.*, 2020, **120**, 1620–1689.
17 X.-Y. Xu, L.-G. Liu, L.-C. Xu, S.-Q. Zhang and X. Hong, *J. Am. Chem. Soc.*, 2025, **147**, 15318–15328.
18 S. Gallarati, E. M. Bucci, A. G. Doyle and M. S. Sigman, *Nature*, 2026, **651**, 637–646.
19 L. C. Gallegos, G. Luchini, P. C. St. John, S. Kim and R. S. Paton, *Acc. Chem. Res.*, 2021, **54**, 827–836.
20 J. C. A. Oliveira, J. Frey, S.-Q. Zhang, L.-C. Xu, X. Li, S.-W. Li, X. Hong and L. Ackermann, *Trends Chem*, 2022, **4**, 863–885.
21 L. Xu, D. Zhu, M. Tang, T. Liu, Y. Liu, C. Ma, Z. Li, S. Zhang and X. Hong, *Chem.–Eur. J.*, 2026, e71074.
22 M. Moskal, W. Beker, S. Szymkuć and B. A. Grzybowski, *Angew. Chem., Int. Ed.*, 2021, **60**, 15230–15235.
23 J. P. Reid, I. O. Betinol and Y. Kuang, *Chem. Commun.*, 2023, **59**, 10711–10721.
24 L.-C. Xu, J. Frey, X. Hou, S.-Q. Zhang, Y.-Y. Li, J. C. A. Oliveira, S.-W. Li, L. Ackermann and X. Hong, *Nat. Synth.*, 2023, **2**, 321–330.
25 S. Zhang, L. Xu, S. Li, J. C. A. Oliveira, X. Li, L. Ackermann and X. Hong, *Chem.–Eur. J.*, 2023, **29**, e202202834.
26 M. E. Ruos, N. P. Romer, J. A. Deichert, L. M. Alabanza, S. S. Gandhi, G. Z. Brown, R. C. Walroth, K. Cruz, F. Gosselin, A. Y. Hong, M. S. Sigman and A. G. Doyle, *J. Am. Chem. Soc.*, 2025, **147**, 25815–25824.
27 D. Dalmau, M. S. Sigman and J. V. Alegre-Requena, *Chem. Sci.*, 2025, **16**, 8555–8560.
28 L.-C. Xu, M.-J. Tang, J. An, F. Cao and Y. Qi, *Nat. Mach. Intell.*, 2025, **7**, 1561–1571.
29 L.-C. Xu, J. An, W. Liu, Y.-F. Shi, C.-L. Ji, F. Cao and Y. Qi, *ChemRxiv*, 2026, preprint, DOI: 10.26434/chemrxiv.10001667/v2.
30 A. Verloop, W. Hoogenstraaten and J. Tipker, in *Drug Design*, Elsevier, 1976, pp. 165–207.
31 L.-C. Xu, X. Li, M.-J. Tang, L.-T. Yuan, J.-Y. Zheng, S.-Q. Zhang and X. Hong, *Synlett*, 2021, **32**, 1837–1842.
32 A. C. Hillier, W. J. Sommer, B. S. Yong, J. L. Petersen, L. Cavallo and S. P. Nolan, *Organometallics*, 2003, **22**, 4322–4326.
33 A. F. Zahrt, J. J. Henle, B. T. Rose, Y. Wang, W. T. Darrow and S. E. Denmark, *Science*, 2019, **363**, eaau5631.
34 A. V. Brethomé, S. P. Fletcher and R. S. Paton, *ACS Catal.*, 2019, **9**, 2313–2323.
35 D. T. Ahneman, J. G. Estrada, S. Lin, S. D. Dreher and A. G. Doyle, *Science*, 2018, **360**, 186–190.
36 X. Li, S. Zhang, L. Xu and X. Hong, *Angew. Chem., Int. Ed.*, 2020, **59**, 13253–13259.
37 K. C. Harper, E. N. Bess and M. S. Sigman, *Nat. Chem.*, 2012, **4**, 366–374.

38 G. S. Tâmega, M. O. Costa, A. De Araujo Pereira and M. A. Barbosa Ferreira, *Chem. Rec.*, 2024, **24**, e202400148.
39 S. Escayola, N. Bahri-Laleh and A. Poater, *Chem. Soc. Rev.*, 2024, **53**, 853–882.
40 B. A. Stenfors, J. A. Cadge, S. Aikonen, G. Luchini, J. Wahlers, K. Koh, M. Muuronen, M. Menche, M. Pfeifle, A. Keto, R. S. Paton, M. S. Sigman and O. Wiest, *J. Org. Chem.*, 2025, **90**, 13874–13884.
41 J. A. Cadge, S. D. Hart, R. C. Walroth, K. A. Mack and M. S. Sigman, *Chem. Sci.*, 2025, **16**, 20473–20485.
42 P. Schwaller, D. Probst, A. C. Vaucher, V. H. Nair, D. Kreutter, T. Laino and J.-L. Reymond, *Nat. Mach. Intell.*, 2021, **3**, 144–152.
43 W. Liu, F. Cao, Y. Qi and L.-C. Xu, *arXiv*, 2026, preprint, arXiv:2601.03689, DOI: 10.48550/arXiv.2601.03689.
44 T. Gensch, G. Dos Passos Gomes, P. Friederich, E. Peters, T. Gaudin, R. Pollice, K. Jorner, A. Nigam, M. Lindner-D'Addario, M. S. Sigman and A. Aspuru-Guzik, *J. Am. Chem. Soc.*, 2022, **144**, 1205–1217.
45 G. Yu, K. Yu, X. Wang, C. Zhang, Y. Luo, X. Huo and Y. Yang, *J. Cheminf.*, 2025, **17**, 45.
46 S. Gallarati, P. Van Gerwen, R. Laplaza, S. Vela, A. Fabrizio and C. Corminboeuf, *Chem. Sci.*, 2022, **13**, 13782–13794.
47 S. Yu, *Chem.–Asian J.*, 2025, **20**, e202500023.
48 M. Tang, T. Zhang, Q. Huang, S. Li, R. Liu, H. Li, X. Chen, S. Dong, X. Liu, X. Feng and X. Hong, *Angew. Chem., Int. Ed.*, 2025, e18560.
49 RDKit: open-source chemoinformatics and machine learning, https://www.rdkit.org.
50 C. Bannwarth, S. Ehlert and S. Grimme, *J. Chem. Theory Comput.*, 2019, **15**, 1652–1671.
51 C. Bannwarth, E. Caldeweyher, S. Ehlert, A. Hansen, P. Pracht, J. Seibert, S. Spicher and S. Grimme, *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, 2021, **11**, e1493.
52 R. Hadsell, S. Chopra and Y. LeCun, in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition – Volume 2 (CVPR'06)*, IEEE, New York, NY, USA, 2006, vol. 2, pp. 1735–1742.
53 A. van den Oord, Y. Li and O. Vinyals, *arXiv*, 2019, preprint, arXiv:1807.03748, DOI: 10.48550/arXiv.1807.03748.
54 R. Shi, G. Yu, X. Huo and Y. Yang, *J. Cheminf.*, 2024, **16**, 22.
55 Y.-L. Liao, B. M. Wood, A. Das and T. E. Smidt, in *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7–11, 2024*, OpenReview.net, 2024.
56 Y. LeCun, S. Chopra, R. Hadsell, M. Ranzato and F. Huang, *Predicting structured data*, 2006.
57 S. Kullback and R. A. Leibler, *Ann. Math. Stat.*, 1951, **22**, 79–86.
58 M. Tokunaga, J. F. Larrow, F. Kakiuchi and E. N. Jacobsen, *Science*, 1997, **277**, 936–938.
59 M. Glos and O. Reiser, *Org. Lett.*, 2000, **2**, 2045–2048.
60 G. Liao, T. Zhang, Z. Lin and B. Shi, *Angew. Chem., Int. Ed.*, 2020, **59**, 19773–19786.
61 R. Connon, B. Roche, B. V. Rokade and P. J. Guiry, *Chem. Rev.*, 2021, **121**, 6373–6521.
62 H. Shimizu, I. Nagasaki and T. Saito, *Tetrahedron*, 2005, **61**, 5405–5432.
63 P.-F. Zhao, K. Wang, J.-X. Wen, Z.-M.-D. Zhu, H. Zhang, Z.-H. Wang, Y.-X. Liao, C.-S. Da and Z.-H. Du, *Org. Biomol. Chem.*, 2025, **23**, 7872–7913.
64 C. A. Tolman, *Chem. Rev.*, 1977, **77**, 313–348.
65 L. van der Maaten and G. Hinton, *J. Mach. Learn. Res.*, 2008, **9**, 2579–2605.
66 C. A. Tolman, *J. Am. Chem. Soc.*, 1970, **92**, 2953–2956.
67 J. Y. Wang, J. M. Stevens, S. K. Kariofillis, M.-J. Tom, D. L. Golden, J. Li, J. E. Tabora, M. Parasram, B. J. Shields, D. N. Primer, B. Hao, D. Del Valle, S. DiSomma, A. Furman, G. G. Zipp, S. Melnikov, J. Paulson and A. G. Doyle, *Nature*, 2024, **626**, 1025–1033.
68 D. Rogers and M. Hahn, *J. Chem. Inf. Model.*, 2010, **50**, 742–754.
69 H. Stärk, D. Beaini, G. Corso, P. Tossou, C. Dallago, S. Günnemann and P. Lió, in *Proceedings of the 39th International Conference on Machine Learning*, ed. K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu and S. Sabato, PMLR, 2022, vol. 162, pp. 20479–20502.
70 X. Ji, Z. Wang, Z. Gao, H. Zheng, L. Zhang, G. Ke and W. E, in *Proceedings of the 38th International Conference on Neural Information Processing Systems*, Curran Associates Inc., Red Hook, NY, USA, 2024.
71 P. Geurts, D. Ernst and L. Wehenkel, *Mach. Learn.*, 2006, **63**, 3–42.
72 H. Drucker, C. J. C. Burges, L. Kaufman, A. Smola and V. Vapnik, in *Advances in Neural Information Processing Systems*, ed. M. C. Mozer, M. Jordan and T. Petsche, MIT Press, 1996, vol. 9.
73 F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and É. Duchesnay, *J. Mach. Learn. Res.*, 2011, **12**, 2825–2830.
74 D. E. Graff, E. I. Shakhnovich and C. W. Coley, *Chem. Sci.*, 2021, **12**, 7866–7881.
75 S. Ma, Y. Cao, Y.-F. Shi, C. Shang, L. He and Z.-P. Liu, *Chem. Sci.*, 2024, **15**, 13359–13368.
76 L. Xu, S. Zhang, X. Li, M. Tang, P. Xie and X. Hong, *Angew. Chem., Int. Ed.*, 2021, **60**, 22804–22811.