Digital Discovery

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal

Cite this: DOI: 10.1039/d6dd00129g

Taming T-REX: a canonical language for geometry-aware generative design of transition-metal complexes

Ilia Kevlishvili * and Devmin Dorabawila

Canonical string representations have transformed organic cheminformatics, yet transition-metal complexes (TMCs) lack an equivalent that captures coordination geometry, stereochemistry, and donor topology. We introduce Trans-pair Relations EXpression (T-REX), a canonical line notation encoding geometry, topology, and metal-centered chirality (@/@@, Δ/Λ) *via trans*-pair maps. Applied to 63 375 DFT-optimized structures from the tmQMg dataset, T-REX identifies five distinct isomer classes (coordination, enantiomeric, linkage, hemilabile, and geometric) and reveals that fewer than 1.2% of complexes capable of stereoisomerism are resolved as such in crystallographic data. Combinatorial enumeration expands these parent structures into 149 228 unique topological variants; modular ligand substitution generates millions of additional candidates. Across one bond-only baseline and four geometry-aware architectures, encoding the T-REX coordination map consistently improves prediction of HOMO, LUMO, gap, and dipole moment. Dipole moment shows the largest gains (R^2 = 0.845 *vs.* 0.715 for the baseline), and three architecturally distinct models with a direct coordination-sphere readout achieve equivalent performance, confirming that T-REX topology, not architecture choice, drives the improvement. Geometry-aware models reach equivalent accuracy with roughly four times less training data, positioning T-REX as both an interoperable data format and an ML-ready representation for transition-metal chemistry.

Received 19th March 2026
Accepted 27th May 2026

DOI: 10.1039/d6dd00129g

rsc.li/digitaldiscovery

1 Introduction

String-based molecular representations have become the default representation of chemical data for main group chemistry.[1] These representations compress structures into compact, human-readable, machine-interpretable text that slots straight into existing cheminformatics[2] and ML workflows.[3–6] Beyond decades of success with SMILES[7] and InChI,[8] newer variants such as DeepSMILES[9] and SELFIES[10] were explicitly designed with modern ML in mind, improving robustness and downstream tokenization, making strings a natural fit for large language model (LLM)-driven optimization and design loops.[11]

Extending this success from organics to transition-metal complexes (TMCs) is nontrivial.[12] Metal complexes span multiple coordination numbers, each with distinct, chemistry-relevant geometries; they carry additional electronic descriptors (oxidation state, multiplicity); and they exhibit a larger number of coordination isomers (*e.g.*, an octahedral complex with 6 unique ligands has 30 stereoisomers, 15 enantiomeric pairs). Haptic and polydentate ligation further complicate topology. At the same time, the surge of string-focused advances (*e.g.*, SELFIES[10,13] for robust small-molecule generation; BigSMILES[14–16] for stochastic polymers) underscores a broad community desire for representations that can carry richer chemical domains, motivating an inorganic-aware string that captures geometry, stereochemistry, and donor identity without sacrificing ML-readiness.

Recent efforts point in this direction but leave key gaps. Rasmussen *et al.* introduced an RDKit-parsable SMILES workflow for TMCs that converts 3D structures into SMILES,[17] improving interoperability but not canonically resolving geometry/symmetry or electronic-state labeling at the string level. In parallel, the automated coordination complex conformer generator, MetalloGen,[18] proposed m-SMILES: an input dialect that encodes the metal, per-ligand strings, explicit coordination sites, and a geometry tag to drive 3D conformer generation, powerful for building, but dependent on non-canonical site numbering. Meanwhile, descriptor families like RACs[19] have been widely used to predict TMC properties.[19–25] Ligand-derived features have commonly been used as a representation, but fail to generalize across different TMCs.[26–30] Geometry/quantum-aware features deliver accuracy but require 3D coordinates[31,32] or QM features,[33] often from DFT, limiting their scale. Furthermore, string representations have been actively used in LLM-driven optimization[34] and structure generation.[35–44] These threads collectively motivate the need for a canonical string-level solution, one where a specific chemical

Department of Chemistry and Biochemistry, Baylor University, Waco, Texas, USA. E-mail: ilia_kevlishvili@baylor.edu; Tel: +1-254-710-4272

species maps to exactly one deterministic string, ensuring database integrity and preventing duplicate bias in machine learning models. Canonical string-level resolution is critical because it makes the representation a chemically meaningful key: identical complexes collapse to one string, while distinct coordination geometries, stereoisomers, linkage modes, oxidation states, and spin states remain separable. This prevents duplicate bias during dataset merging and avoids conflating isomeric species in enumeration or ML workflows.

In this work, we (i) introduce Trans-pair Relations EXpression (T-REX), a canonical line notation for monometallic CN $\leq$ 7 complexes that encodes geometry, coordination topology, and metal-centered chirality (@/@@, Δ/Λ) *via trans*-pair maps and stereochemical flags; (ii) develop a structure to string extraction pipeline that converts over 63 000 literature structures into canonical T-REX strings and classifies their isomer relationships across five distinct categories; (iii) show how these strings enable systematic enumeration of coordination isomers and enantiomers, as well as ligand-substitution neighborhoods, yielding hundreds of thousands of topological variants and millions of chemically plausible complexes; and (iv) demonstrate across five neural network architectures that T-REX-derived coordination topology consistently improves property predictions, with the largest gains on shape-sensitive properties like dipole moment, and that a direct coordination-sphere readout provides a roughly four-fold improvement in data efficiency over bond-only baselines. In contrast to prior TMC string dialects that prioritize structure generation or RDKit interoperability, T-REX is designed from the outset to be canonical, geometry-aware, and ML-ready at the string level.

2 Syntax & canonicalization

2.1 Syntax

The design philosophy behind the string representation was motivated by a need for a format that is both chemically intuitive and easily human-readable, yet rigorous enough to support canonicalization. A primary objective was to resolve the issue of topological isomers collapsing into identical representations while retaining structural simplicity. We approached this by modeling transition metal and organometallic complexes through the lens of traditional chemical understanding, viewing them as a central metal surrounded by ligands defined by their relative orientations. We show below that by exclusively capturing which ligand coordinating atoms are *trans* to one another and which have no *trans* partner, we can successfully disambiguate not only coordination topology but also coordination geometry. For example, in an octahedral complex with two ligands, A and B, a ***fac*** topology is indicated if all *trans* pairs are A/B, whereas a ***mer*** topology contains A/A, A/B, and B/B pairs. Additionally, for a complex with four ligand sites, the presence of two *trans* pairs implies square planar geometry, a single *trans* pair with two singles implies seesaw geometry, and the total absence of *trans* pairs implies tetrahedral geometry (Fig. 1, SI S1–S7).

T-REX is a modular, line-based notation composed of separable blocks, each delimited by a vertical bar | to make parsing trivial. The header encodes the central transition metal, its oxidation state, and an optional spin multiplicity (default is interpreted as multiplicity = 1 if omitted). Next, the ligand block lists every coordinated ligand's identity. The third block is the map, which records which coordinating atoms (catoms) are *trans* to one another (pairs) and which donors have no *trans* partner (singles). Two optional blocks may follow: a geometry flag (to make the intended idealized CN geometry explicit when helpful) and a central-chirality flag (to disambiguate metal-centered or Δ/Λ stereochemistry). In short, the representation has the following structure:

METAL | LIGANDS | MAP [| GEOM] [| CHIRAL]

The header begins with the element symbol and encloses electronic state in curly braces: `Pd{+2}` or `Fe{+2,5}` (the latter explicitly sets multiplicity M; if M is absent, multiplicity defaults to 1). Oxidation state is mandatory; whereas multiplicity implicitly defaults to a closed-shell species. This compact header keeps the electronic specification orthogonal to topology, so downstream tools can read or ignore it without touching connectivity.

The ligand list is introduced by L = and wrapped in square brackets: `L=[lig1, lig2, … ]`. Each ligand is preceded by a payload tag that declares how to interpret its string, enabling modular growth of the string representation. For example, `SMILES:` supports RDKit-centric workflows and structure generation; a future extension to `SELFIES:` is a natural representation for generative models; a future semantic payload can maximize human readability. The current software relies on `SMILES:` and separates ligands by commas, *e.g.* `L=[SMILES: [Cl-], SMILES: O=C(C)C(=O)[O-], SMILES: c1ccccn1]`. Donor atoms ("catoms") are indexed relative to each individual ligand string (1-based) and referenced in the map. The ligand payload itself can be changed while updating coordinating atom indices in the map, without breaking the T-REX string.

The map captures the local metal topology using only *trans* pairs and singles, which is sufficient to disambiguate the vast

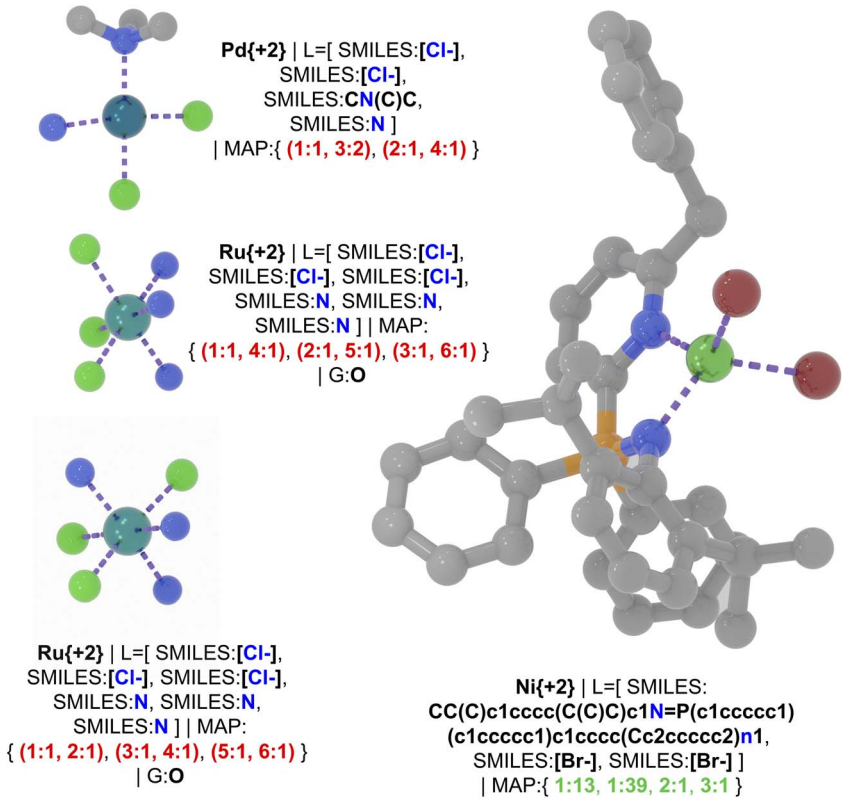

Fig. 1 Examples of T-REX strings for square planar (top left), tetrahedral (right), *fac*-octahedral, and *mer*-octahedral complexes. Coordinating atoms are highlighted in blue, *trans*-pairs are highlighted in red, "singles" are highlighted in green.

majority of coordination geometries and all *cis* relationships up to CN ≤ 6 (Fig. 1). Simultaneously, the map reveals ligand denticity (multiple catoms from the same ligand) and hapticity (groups of catoms treated as a single coordination site). The block begins with `MAP:{...}`. A *trans* pair is written as `(i:a, j:b)`, where `i` and `j` are ligand indices (from the ligand list order) and a, b are their catom indices (from each ligand's string), all 1-based. For example, `(1:11, 2:4)` means ligand 1 atom 11 is *trans* to ligand 2 atom 4. Singles (donors without a *trans* partner) are written as `(i:a)` entries. Haptic donors are grouped, *e.g.* `(1:[2,3], 2:1)` indicates ligand 1's atoms 2 and 3 act as a single haptic site (η^2) and are *trans* to ligand 2 atom 1.

Although the pair/single map alone typically fixes geometry for CN ≤ 6, an explicit geometry flag (*e.g.*, G:O, G:TP) can be appended to guard against rare edge cases (in particular, when clarifying intended CN = 6 geometry families and avoiding accidental conflation of octahedral geometry with less common alternatives like trigonal prismatic geometry). A central chirality flag further locks in metal-center stereochemistry when two enantiomeric assignments share the same pair pattern. T-REX distinguishes two mechanistically distinct types of metal-centered chirality. Point-central chirality (@/@@) is computed *via* the determinant method: four sites are selected according to geometry-specific rules, and the sign of their scalar triple product assigns handedness. Achirality is detected before computation through geometry-specific checks. For instance, equivalent *trans* partners within a pair or equivalent pair sets in octahedral complexes preclude chirality. In contrast, equivalent sites that do not share such relationships do not (full conditions for each geometry are given in SI, Text S5). Helical chirality (Δ/Λ) arises in octahedral complexes bearing multidentate ligands: tris-bidentate, *cis*-bis-bidentate, and *fac–fac* bis-tridentate, where point-central chirality is absent, but a propeller-like twist exists. The chirality flag is assigned during structure-to-string conversion and is preserved through canonicalization. We check the point-central chirality first, then fall back to helical chirality. While enantiomers exhibit identical scalar properties in achiral environments, resolving them at the string level is essential for applications in asymmetric catalysis and biological recognition, and ensures that each physically distinct species maps to a unique T-REX string (Fig. 2).

Importantly, T-REX strings do not need to be extracted from 3D structures. The modular block design allows direct construction from chemical intent where either a user or generative algorithm specifies the metal, oxidation state, ligand set, and desired coordination map, enabling bottom-up dataset construction for hypothetical complexes that have never been synthesized or computationally optimized.

T-REX is designed for CN ≤ 7 where *trans*-pair semantics cleanly characterize geometry and coordination isomerism; nothing in the syntax forbids higher CN, but complete conformer disambiguation may require additional relations beyond *trans* pair enumeration. For example, to avoid the

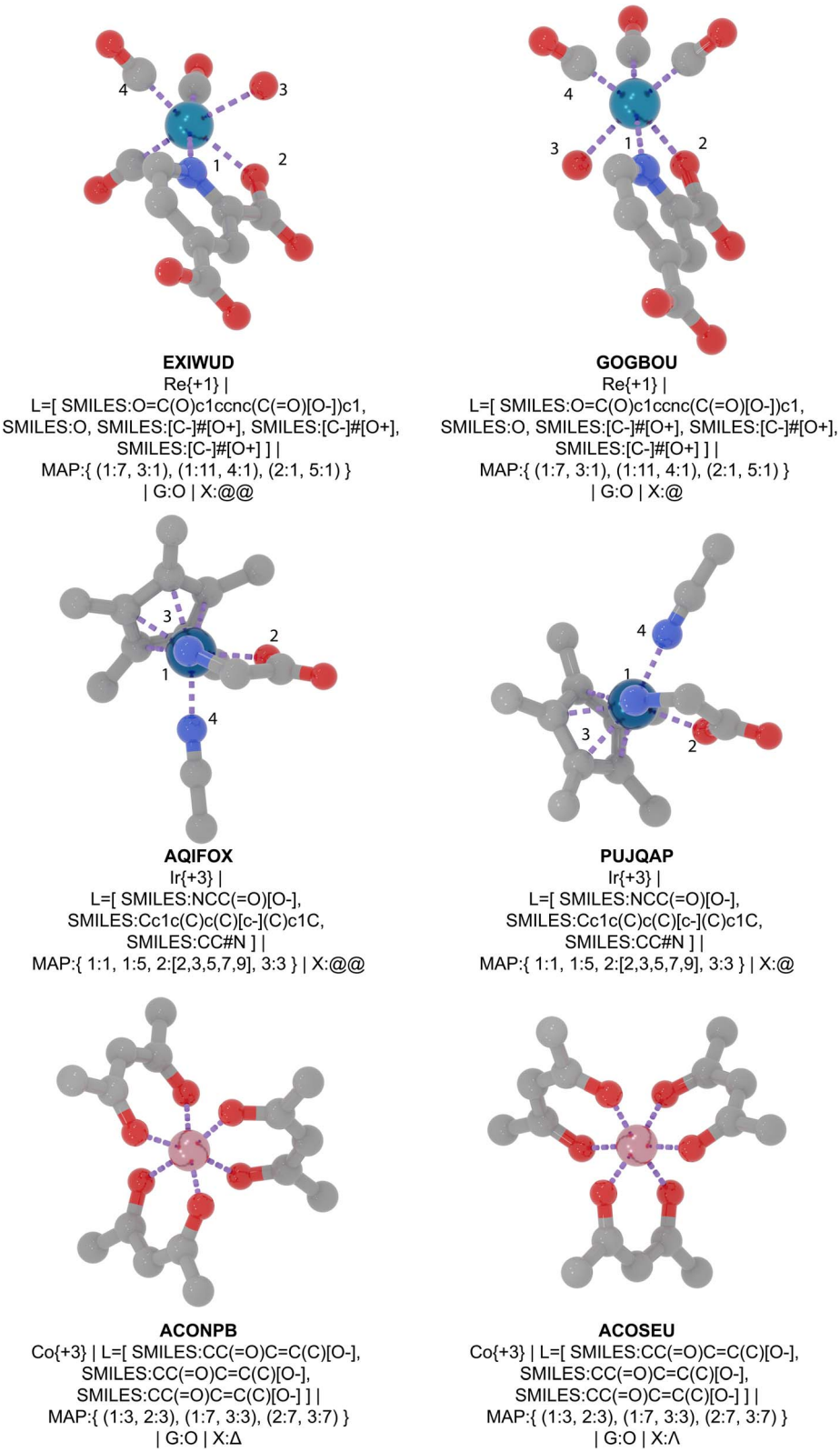

Fig. 2 Examples of chiral molecules represented in T-REX. Atoms used to compute point-central chirality are labeled.

collapse of pentagonal bipyramidal geometry (CN = 7) stereoisomers, equatorial ligands (singles in T-REX) are listed in counterclockwise direction, when looking down from the first ligand in the axial position. T-REX also intentionally discretizes continuous coordination environments into idealized topology classes. Therefore, complexes with the same donor set, geometry label, and *trans*-pair map but different bond lengths, angular distortions, or Jahn–Teller elongation/compression

map to the same T-REX topology unless the distortion changes the assigned coordination geometry or *trans*-pair relationships. This makes T-REX appropriate for canonical topology, isomer enumeration, dataset curation, and geometry-aware ML, but not a replacement for 3D coordinates when quantitative distortion amplitudes are required. Future work will focus on addressing these limitations and expanding grammar for the multinuclear MUL-T-REX, which generalizes the same block structure while preserving the canonical, edit-friendly design.

Extending T-REX to multinuclear complexes introduces several algorithmic challenges beyond simply allowing multiple metal headers. A multinuclear representation must preserve each metal's local coordination topology while also encoding intermetal relationships, including metal–metal bonds and bridging ligands that participate in more than one coordination sphere. This makes canonicalization more complex because both ligand order and metal-center order can permute, and equivalent metal centers must be recognized without changing the local maps. A natural MUL-T-REX extension would therefore treat a cluster as a graph of local T-REX-like metal environments connected by shared ligand sites and explicit metal–metal edges. This preserves the same canonical, edit-friendly philosophy while adding the additional bookkeeping required for clusters, bioinorganic cofactors, and multinuclear catalysts.

2.2 Canonicalization

For a string representation to serve as a database key or an ML input, it must be canonical: a specific chemical species must map to exactly one string. T-REX achieves this through a strict, hierarchical sorting algorithm applied after the generation of a non-canonical, valid T-REX string:

(1) Ligand Sorting: ligands are first canonicalized individually (using RDKit standard canonicalization for SMILES payloads) and coordinating atom indices are remapped to the updated SMILES string. They are then sorted within the Ligand Block based on a priority rule set of decreasing denticity, hapticity, coordinating atom atomic number, ligand molecular weight and increasing ligand hash, in that order.

(2) Map Minimization: once ligand order is fixed, the topology map is sorted to minimize the numerical indices of the *trans*-pairs and singles (lexicographic sorting). For polydentate ligands with internal symmetry, canonicalization ensures that equivalent donor permutations map to a single T-REX string.

For example, in a square planar $Pd(Cl)_2(NH_3)_2$ complex, the T-REX algorithm ensures that the chloride ligands (higher atomic number) are always listed before amines, and the map is ordered such that *cis* and *trans* isomers yield deterministic, non-overlapping strings. This guarantees that T-REX is invariant to atom indexing in the source file.

3 Results and discussion

3.1 T-REX datasets

We validated the representation's coverage by converting the tmQMg dataset,[33] which is a comprehensive library of DFT-optimized transition-metal complex geometries, into canonical T-REX strings. The tmQMg dataset consists of structures optimized at the PBE-D3BJ/def2-SVP level with the calculated properties derived from a single point energy calculation at the PBE0-D3BJ/def2-TZVP level of theory. All structures are constrained to the total molecular charge between −1 and 1, with even electrons, and all properties computed as a closed shell singlet. Of the 74 547 starting structures, the conversion pipeline successfully parsed 72 733 (97.5%), with rejections arising from charge-assignment parser failures (1,661) and processing timeouts (153). Parsed structures were then subjected to a geometry-agreement filter in which the coordination geometry inferred independently from the T-REX *trans*-pair/singleton pattern was compared against the molSimplify RMSD-based classification; only complexes where both methods agreed were retained, yielding 66 525 structures. Given that tmQMg reports properties for closed-shell singlet electronic structures, the same electronic-state assignment was used for the T-REX dataset and ML benchmarks. Although T-REX can encode spin multiplicity explicitly, the present benchmarks do not evaluate spin-state ordering, spin-crossover energetics, or open-shell alternatives. A multi-step cleanup further refined this set: 94 entries with erroneous explicit-hydrogen placement were corrected, element counts in each T-REX string were audited against the source *XYZ* file to detect ligation-state misassignments, and 57 structures were rescued by identifying cases where the extended Hückel charge-assignment workflow had misinterpreted perchlorate ligands as oxo species. After removing the remaining 2007 atom-count mismatches and applying canonical deduplication (1143 duplicates), the final dataset comprised 63 375 unique T-REX strings spanning all major coordination geometries, establishing T-REX as a robust format for large-scale curation of inorganic and organometallic data. Additionally, the same pipeline was applied to four previously published functional datasets derived from tmQMg[45] (tmCAT, tmPHOTO, tmBIO, and tmSCO), yielding 18 855, 4,061, 2,542, and 8209 unique complexes, respectively.

The curated tmQMg library spans 52 transition metals across oxidation states from −3 to +7 (SI Fig. S8), with Pd (7,194), Pt (5,832), Ru (5,501), Ni (5,211), and Zn (5,058) as the most represented. Because T-REX jointly encodes metal identity, oxidation state, and coordination geometry, the dataset enables direct quantification of metal–geometry coupling. Several metals exhibit strong geometric preferences. For example, Pd is 94% square planar and Au is 87% linear, while others display increased diversity. Ru splits nearly evenly between tetrahedral (51%) and octahedral (40%), and Zn populates multiple distinct geometry families (SI Fig. S9). These distributions reflect the complexity of transition metal complexes that T-REX captures at scale.

3.2 Enumeration of coordination isomers

We performed systematic isomer classification across the 63 375 unique tmQMg complexes, revealing five distinct classes of isomer relationships among structures sharing the same metal and ligand set. Coordination isomers with complexes differing

only in their *trans*-pair map were the most prevalent, with 254 sets (516 structures, including sets of up to three resolved diastereomers). Enantiomeric pairs resolved by the chirality flag accounted for 92 sets (184 structures), while linkage isomers, in which the same ligand coordinates through a different donor atom, comprised 52 sets (104 structures). Hemilabile isomers, in which a multidentate ligand partially dissociates to change its effective denticity, accounted for 11 sets (22 structures). It is noteworthy that while several hemilabile ligands have been identified in the CSD in past work,[46–48] a very limited number are characterized as hemilabile complexes in the same coordination environment. Geometric isomers, where the same composition adopts an entirely different coordination geometry, appeared as 8 sets (16 structures). Two additional sets flagged as identical (4 structures) are pentagonal bipyramidal complexes where the current canonicalization is surjective but not injective, confirming that the representation is otherwise bijective across all supported geometry families. Representative examples of enantiomeric pairs and coordination isomers are shown in Fig. 2 and 3, respectively, and other examples are shown in SI Fig. S10–S12.

Notably, of the 63 375 unique complexes, 29 491 (46.5%) are theoretically capable of coordination isomerism or enantiomerism, yet only 338 unique sets contained resolved coordination isomers or enantiomeric pairs; 8 of these are triplets in which a coordination isomer pair is accompanied by a resolved enantiomer within one of the diastereomeric forms (*e.g.*, $Co(III)(en)_2(N_3)_2$, Fig. 3). This scarcity systematically obscures the geometric and stereochemical contrast necessary for ML models to learn isomer-dependent properties for transition metal complexes. By applying a combinatorial enumeration algorithm that permutes *trans*-pair and singleton assignments while preserving multidentate constraints, and assigns both enantiomeric forms where chirality is present, we expanded the 63 375 parent structures into 149 228 unique canonical T-REX strings, capturing the topological diversity that crystallographic databases leave unresolved. Separately, the 12 370 complexes identified as chiral during the conversion were reflected to generate their mirror-image enantiomers, yielding 12 370 paired structures with explicit chirality labels (@/@@, Δ/Λ) and DFT-quality geometries. This enantiomer library[49] is provided as a standalone dataset for applications in asymmetric catalysis and bioinorganic design (T-REX-ent).

3.3 Generative expansion using chemically and topologically plausible ligand substitutions

The modular architecture of T-REX facilitates the generation of massive combinatorial libraries by treating ligands as interchangeable components within a fixed topology. By classifying ligands into "strict" (same donor atom identity) and "relaxed" (scaffold hopping) compatibility groups, we enable substitutions that preserve the overall geometry of the complex while exploring new chemical neighborhoods. For instance, in the tmCAT dataset,[45] this approach condensed over 13 000 unique ligands into just 185 relaxed or 722 strict classes, providing a structured, data-driven basis for combinatorial design (SI Table S1).

To demonstrate this utility for large-scale discovery, we focused on metal hydride complexes within tmCAT, expanding a small parent set into millions of candidates. Starting from just 658 structures containing a metal–H bond, strict substitution generated approximately 717 000 unique canonical strings, while the relaxed approach yielded over 2.3 million unique strings. This massive expansion, spanning several orders of magnitude, confirms that T-REX can rapidly populate the "near" and "far" neighborhoods of synthetically plausible complexes using only valid ligand components (SI Fig. S13).

Furthermore, the integration of SMARTS logic allows for the imposition of specific chemical rules during generation, as

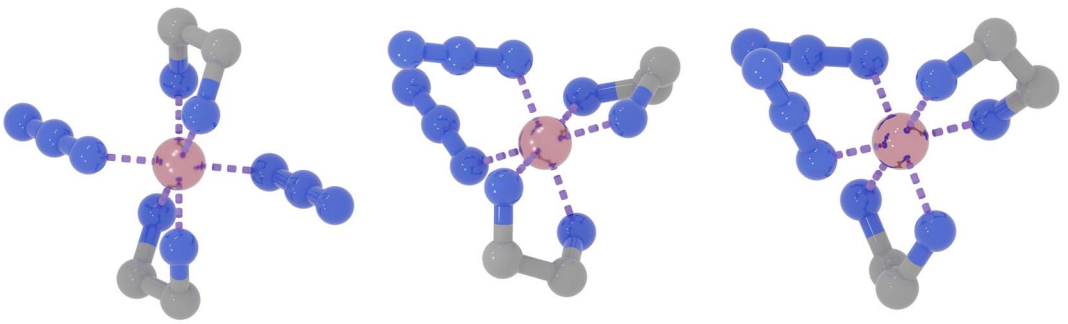

Fig. 3 An example of a coordination isomer triplet found in the tmQMg dataset. Associated CSD REFCODE is shown in black. T-REX string components that are identical are shown in blue, and a distinct component (MAP, chirality block) is shown in green. Associated tmSMILES for these complexes that collapse to the same string representation is shown in red.

demonstrated on cisplatin analogs in the tmBIO dataset. By restricting substitutions to maintain a *cis*-N motif on a subset of 60 parents, we generated nearly 20 000 strict and 25 000 relaxed variants that effectively bridge the chemical space between distinct clusters in the parent dataset (Fig. 4). We extended this workflow to the full tmSCO dataset, where 819 spin-crossover complexes were expanded into a library of ∼160 000 unique T-REX strings, illustrating the method's generalizability across diverse inorganic domains. (SI Fig. S14).

These results demonstrate that ligand classification and substitution enumeration *via* T-REX can drive the generative design of massive combinatorial datasets. Furthermore, the integration of SMARTS logic allows for the imposition of desirable chemical rules during generation. We defined two different ligand classification approaches, with "strict" classification envisioned as a more appropriate tool for local optimization and "relaxed" classification more appropriate for scaffold hopping and discovery. We envision that this modularity will enable efficient genetic algorithm (GA) optimization strategies, where high-level T-REX information defines "genes" for metal topology and electronic structure, while ligands serve as modular "subgenes" for local optimization.

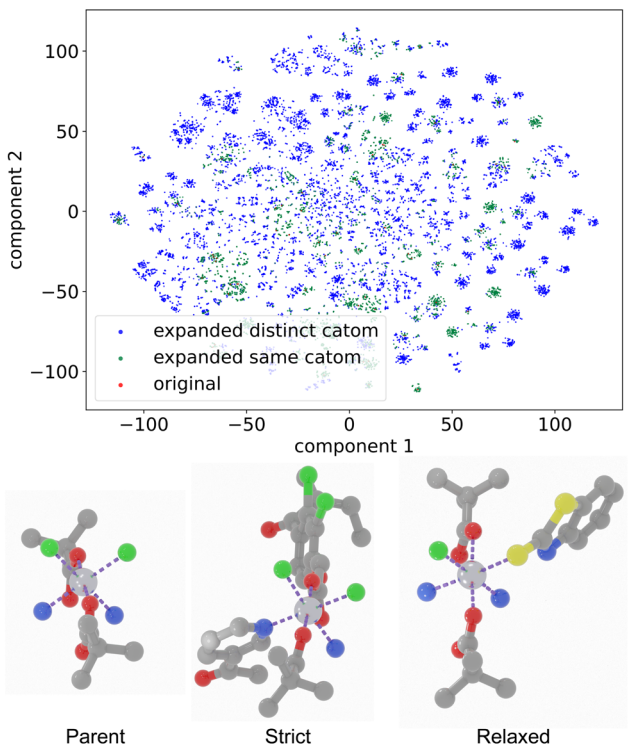

Fig. 4 t-SNE visualization of the chemical space covered by the generated *cis*-diamino complex strings in tmBIO dataset (top), and Pt (IV) structures from these datasets (bottom). Original parent complexes are shown in red, strict expansion is shown in green, and relaxed expansion is shown in blue. The visualized parent compound is *cis*-platin in the equatorial plane with two axial carboxylate ligands. The strict di-substitution shows replacement of the ammonia with pyridine-type ligand and one carboxylate. The relaxed mono-substitution leads to the replacement of chloride with thiolate ligand. Representative 3D depictions of selected strict and relaxed products were constructed from the corresponding T-REX strings and optimized with GFN2-xTB for visualization only.

3.4 Graph and hypergraph neural networks

To validate T-REX as a machine-learning-ready format, we integrated strings into five graph neural network architectures[50] and evaluated their ability to predict DFT-calculated electronic properties directly from the 2D graph, without any 3D coordinates as input. We utilized the T-REX-converted tmQMg dataset to predict four properties obtained from the tmQMg[33] DFT labels: HOMO energy, LUMO energy, HOMO–LUMO gap, and dipole moment. We used a random 80/10/10 split, with 50 700 complexes in the training set and 6337 complexes in each of the validation and test sets. For each architecture, we performed hyperparameter optimization using Optuna and report ensemble predictions averaged over five independently seeded runs.

We contrasted a bond-only baseline, which ablates hyperedge message passing, but retains the full molecular graph, against four geometry-aware architectures that differ in how they process and route T-REX-derived coordination topology. The baseline Message Passing Neural Network (MPNN) utilizes the GINE convolutional architecture,[51] with node features including one-hot element encodings, RDKit-derived properties, Pauling electronegativity, and chirality tags, while edge features consist of standard bond types. This model captures bond topology but is effectively blind to stereochemical relationships like *cis*/*trans* isomerism. The four geometry-aware architectures augment this bond graph with hyperedges[52–55] constructed directly from the T-REX *trans*-pair map, where each hyperedge connects two coordination sites (A and B) through the metal center (M), labeled as *cis* or *trans* with a discrete ideal-angle class. They differ along two architectural axes. First is the hyperedge processing mechanism which includes attention-based GRU pooling[56] (HyperMPNN and LF-GNN), DeepSets-style permutation-invariant aggregation[57] (DeepSets), and absorption into virtual graph nodes[58] processed by standard message passing (Virtual Node). Second is the readout pathway, where HyperMPNN routes hyperedge information exclusively through atom features before pooling, whereas LF-GNN, DeepSets, and Virtual Node architectures maintain a direct hyperedge-to-head channel that concatenates atom-level and coordination-level pooled representations (SI, Text S7).

For frontier orbital energies, all geometry-aware architectures proved highly effective and tightly clustered. HOMO prediction yielded R^2 values of 0.980 (LF-GNN), 0.979 (DeepSets), 0.978 (Virtual Node), and 0.977 (HyperMPNN), compared to 0.968 for the bond-only MPNN (SI Fig. S15–S24). LUMO predictions followed a similar pattern, with geometry-aware models spanning R^2 0.972–0.978 *versus* 0.966 for the baseline (SI Fig. S25–S34). The HOMO–LUMO gap, while still largely dictated by ligand-field strength, showed a wider spread: R^2 values ranged from 0.898 to 0.903 for the direct-readout architectures, 0.884 for HyperMPNN, and 0.868 for the baseline (Table 1, SI Fig. S35–S44).

Table 1 Test-set prediction performance (R^2 and MAE) for four DFT-calculated properties across five GNN architectures, reported as ensemble predictions averaged over five independently seeded runs. MPNN is the bond-only baseline; HyperMPNN, LF-GNN, DeepSets, and virtual node incorporate T-REX-derived *cis*/*trans* hyperedges, with the latter three maintaining a direct coordination-sphere readout channel. MAE is reported in eV for HOMO, LUMO, and HOMO–LUMO gap, and in Debye for dipole moment

Property	Dipole (MAE/R^2)	HL gap (MAE/R^2)	HOMO (MAE/R^2)	LUMO (MAE/R^2)
MPNN	1.272/0.715	0.212/0.868	0.160/0.968	0.160/0.966
HyperMPNN	1.097/0.813	0.191/0.884	0.133/0.977	0.142/0.972
LF-GNN	**0.968/0.845**	0.177/0.898	**0.125/0.980**	**0.127/0.978**
DeepSets	0.979/0.843	0.173/0.902	0.131/0.979	0.132/0.976
Virtual node	1.002/0.837	**0.173/0.903**	0.132/0.978	0.130/0.977

However, a stark performance hierarchy emerged when predicting the dipole moment, a vector property intrinsically sensitive to the spatial arrangement of ligands around the metal center. The bond-only MPNN, which treats coordination isomers as identical, achieved $R^2 = 0.715$ (MAE = 1.27 D). HyperMPNN, which encodes T-REX-derived *cis*/*trans* hyperedges but routes their information through atom features, improved substantially to $R^2 = 0.813$ (MAE = 1.10 D). The three architectures with a direct coordination-sphere readout channel performed best and were effectively interchangeable: LF-GNN ($R^2 = 0.845$, MAE = 0.97 D), DeepSets ($R^2 = 0.843$, MAE = 0.98 D), and virtual node ($R^2 = 0.837$, MAE = 1.00 D) (Fig. 5, SI S45–S54 and Table 1). The consistency across three fundamentally different hyperedge processors indicates that the T-REX topology itself, rather than the specific neural architecture, is the primary factor improving the performance of shape-sensitive properties.

Notably, the performance hierarchy reveals that two architectural choices matter independently: encoding coordination topology at all (MPNN to HyperMPNN, $\Delta R^2 \approx 0.10$ on dipole), and providing that topology a direct path to the prediction head (HyperMPNN to LF-GNN, $\Delta R^2 \approx 0.03$ on dipole). The benefit of a dedicated coordination-sphere readout channel parallels the established advantage of separating metal-centered features[19] (mc-RAC) from full-complex descriptors, here realized as learnable two-body representations derived from the T-REX *trans*-pair map. Crucially, all five architectures operate on identical atom and bond featurization. The only variable is whether and how the model accesses the coordination map encoded in the T-REX string, confirming that *trans*-pair encoding at the string level is sufficient to recover the geometry dependence of strongly shape-sensitive properties without explicit 3D coordinates.[59–61]

As a standard cheminformatics comparison, we also trained ECFP4/random-forest models from T-REX-derived RDKit molecular graphs with and without T-REX-derived virtual *trans* bonds. For HOMO–LUMO gap prediction, the bond-only ECFP4/RF model gave $R^2 = 0.644$, close to previously reported RDKit-SMILES fingerprint baselines,[17] while adding virtual *trans* bonds gave $R^2 = 0.625$. For dipole moment, the same comparison improved from $R^2 \approx 0.51$ without *trans* bonds to $R^2 = 0.582$ with *trans* bonds. These results show that T-REX-derived *trans*-pair topology can benefit even classical fingerprints for geometry-sensitive properties, while also confirming that the bond-only MPNN is a stronger learned graph baseline than ECFP/RF.

To assess data efficiency, we trained ensembles of three models at 10%, 25%, 50%, 75%, and 90% of the training data and evaluated on the full test set for all four properties (Fig. 6, SI S55–S57). The magnitude and onset of the geometry-aware advantage scaled directly with the shape-sensitivity of the target property. For dipole moment, the three direct-readout architectures at 25% of the training data (~12 700 complexes) already exceeded the bond-only MPNN trained on the full dataset ($R^2 \approx 0.73$ *vs.* 0.72), representing a roughly four-fold reduction in labeled data needed to reach equivalent accuracy. For the HOMO–LUMO gap, the geometry-aware advantage was consistent but diminished, with clear separation emerging by 25% of training data. For HOMO and LUMO energies, all architectures converged rapidly and showed minimal separation across data fractions, consistent with the weaker geometry dependence of frontier orbital energies, though geometry-aware models maintained a small but consistent edge at full training set size across all properties.

Finally, we tested isomer-resolved prediction using a stricter isomer-holdout split in which all 516 coordination-isomer structures identified in tmQMg were assigned to the test set, while all remaining complexes were split into train and validation sets. This split prevents memorization of family-specific isomer offsets and instead tests whether models can generalize topology-property relationships across chemically distinct isomer families. For dipole prediction, the bond-only MPNN performed poorly on this diagnostic, giving $R^2 = -0.139$ and MAE = 3.03 D, whereas LF-GNN achieved $R^2 = 0.871$ and MAE = 1.16 D (Fig. 7). Pairwise Δ-dipole prediction within isomer families showed an even larger separation: MPNN gave $R^2 = 0.111$ with the majority of pairwise contrasts compressed toward $\Delta = 0$, whereas LF-GNN gave $R^2 = 0.918$ and recovered the direction of large isomer effects (Fig. 7). The MPNN does not collapse every contrast because the ablated graph still includes a metal-centered chirality/achirality feature, which provides symmetry information. For HOMO–LUMO gap, both models retained strong absolute performance on the isomer-holdout set, but within-family Δ-gap prediction was more subtle; LF-GNN improved pairwise Δ-gap R^2 from 0.090 to 0.237 and Spearman correlation from 0.175 to 0.559 (SI Fig. 58). These results show that T-REX-derived coordination topology is most critical for strongly geometry-sensitive targets such as dipole

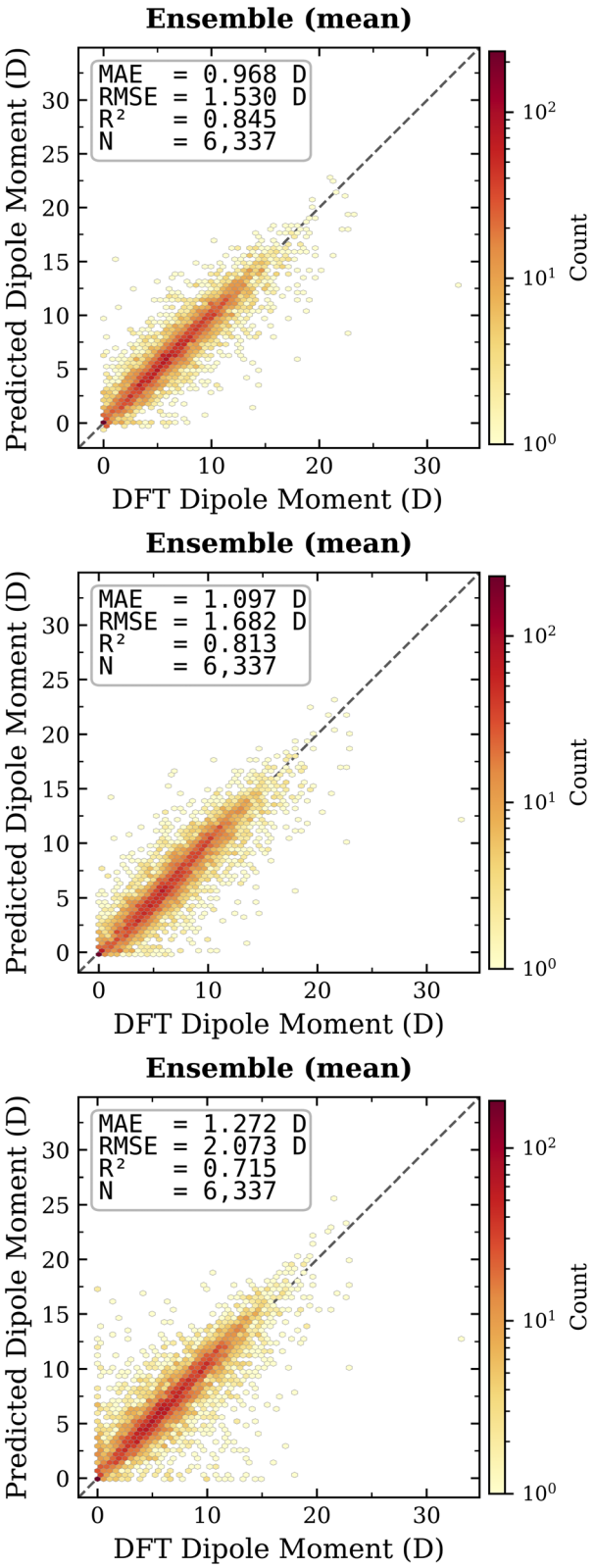

Fig. 5 Parity plots of predicted *vs.* calculated dipole moment on the set aside test set using the LF-GNN (top), HyperMPNN (middle), and MPNN (bottom) architectures.

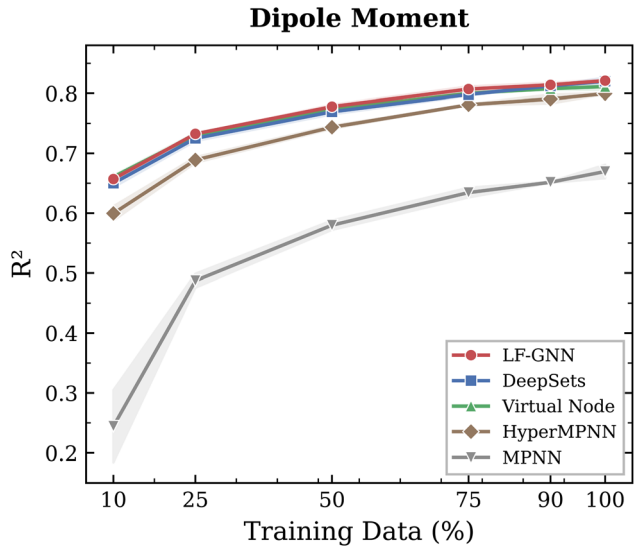

Fig. 6 Learning curves for dipole moment prediction (R^2 *vs.* training data fraction) across five GNN architectures. Shaded regions indicate ±1 standard deviation over three independently seeded runs (five seeds at 100%). The three architectures with a direct coordination-sphere readout (LF-GNN, DeepSets, virtual node) cluster tightly across all data fractions and exceed the bond-only MPNN trained on the full dataset by 25% of training data, while HyperMPNN follows a parallel but consistently lower trajectory.

moment, while more subtle isomer-dependent orbital-energy shifts likely require larger, deliberately completed isomer-resolved training datasets.

4 Outlook and conclusions

Despite its breadth, the present work leaves several opportunities for expansion. First, while T-REX supports CN ≤ 7 with full disambiguation for all major geometry families and metal-centered chirality (@/@@, Δ/Λ), extending the grammar and canonicalization rules to multinuclear architectures (MUL-T-REX) will be essential for covering bioinorganic clusters and heterogeneous motifs. Second, the generative libraries constructed here are combinatorial and "chemically plausible" by design but are not yet filtered by thermodynamic stability, kinetic accessibility, or synthetic feasibility. T-REX can also provide a framework for candidate prioritization. The representation itself is not a synthetic-accessibility or thermodynamic-stability score, but its explicit metal, oxidation-state, ligand, donor-site, geometry, and *trans*-pair fields make it straightforward to apply hierarchical filters. At the string level, inexpensive rules can enforce charge/electron-count constraints, allowed donor–metal combinations, known ligand classes, SMARTS-defined motifs, and user-specified coordination patterns. Surviving candidates can then be converted to 3D structures and ranked using semiempirical or DFT stability estimates, ligand-dissociation energies, spin-state checks, or learned surrogate models that use T-REX-derived hypergraphs for prediction. Thus, T-REX provides a canonical candidate-generation and a framework onto which synthetic-

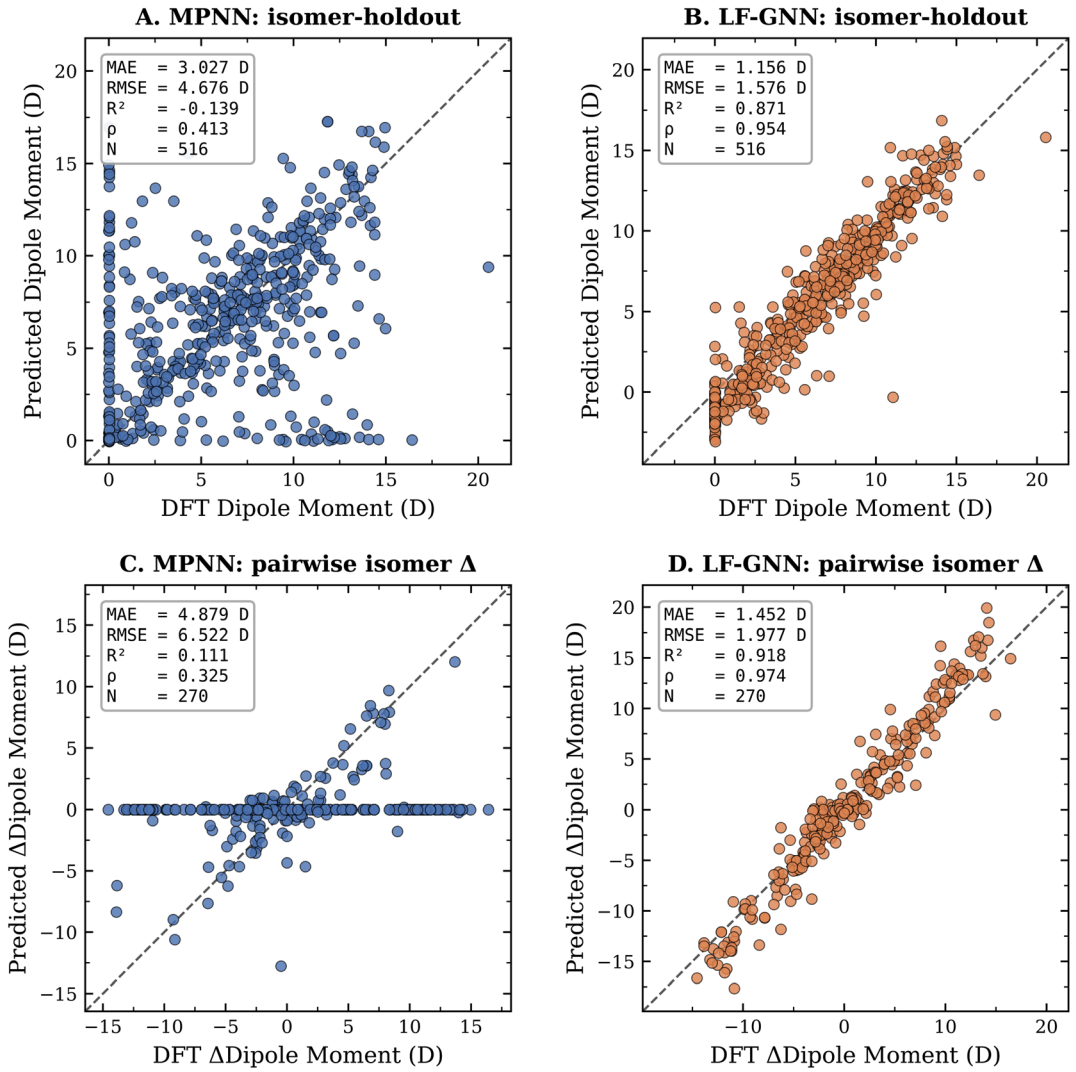

Fig. 7 Isomer-holdout evaluation of T-REX-derived coordination topology. All coordination-isomer families identified in tmQMg were placed in the test set, while the remaining complexes were used for train/validation splitting. Absolute dipole prediction on this held-out set is shown for the bond-only MPNN baseline and LF-GNN in panels (A) and (B), respectively. Panels (C) and (D) show pairwise Δ-dipole prediction for all unordered comparisons within each held-out isomer family. The bond-only MPNN under-resolves many isomer-dependent dipole differences, whereas LF-GNN recovers both absolute dipoles and within-family pairwise contrasts, demonstrating that explicit T-REX coordination topology improves isomer-resolved prediction.

accessibility filters and stability models can be attached. Coupling T-REX-based enumeration with rapid screening (*e.g.*, semiempirical QM, trained surrogates, or expert-encoded SMARTS constraints) and feedback-driven optimization (genetic algorithms, active learning, or reinforcement learning) will be crucial for turning these libraries into practical proposal sets. Finally, while the five GNN architectures benchmarked here demonstrate that T-REX topology provides a strong inductive bias for shape-sensitive properties, scaling to self-supervised or foundation-model regimes on millions of T-REX strings, potentially integrating the enumerated isomer and enantiomer libraries as pretraining data, could unlock broadly transferable representations for TMC catalysis, photophysics, and medicinal inorganic chemistry.

We have introduced T-REX, a canonical line notation that encodes transition-metal complexes as modular strings combining metal identity, electronic state, ligand payloads, a *trans*-pair map, and a metal-centered chirality flag that together uniquely specify coordination topology and stereochemistry for monometallic CN $\leq$ 7 species. An extraction pipeline converts over 63 000 literature structures from the tmQMg dataset into canonical strings, and systematic isomer classification reveals five distinct classes of structural relationships including coordination isomers, enantiomers, linkage isomers, hemilabile isomers, and geometric isomers, while confirming that crystallographic databases dramatically underrepresent the full space of accessible topological variants. By treating T-REX strings as both compact keys and manipulable objects, we enumerate 149 228 unique coordination

isomers and enantiomers, construct large libraries of chemically plausible complexes *via* ligand-class substitutions, and generate a dedicated enantiomer dataset (T-REX-ent) from the 12 370 chiral complexes identified during conversion. Interfacing T-REX with RDKit enables information-enriched graphs and hypergraphs: across five neural network architectures, we show that encoding T-REX-derived coordination topology consistently improves predictions of calculated properties, with the largest gains on dipole moment ($R^2 = 0.845$ *vs.* 0.715 for bond-only baselines), and that this advantage persists at reduced training set sizes, reflecting a roughly four-fold improvement in data efficiency. Together, these results position T-REX as both an interoperable data format and an ML-ready representation for transition-metal chemistry, providing a foundation for more systematic dataset curation, geometry-aware learning, and generative design across catalysis, materials, and bioinorganic discovery.

5 Methods

5.1 General software model

All T-REX functionality is implemented in Python, using Pydantic for schema validation and a modular data model that separates the electronic state (metal, oxidation state, optional spin multiplicity), ligand payloads, and a coordination map over "sites" (ligand index + coordinating atom indices). This object model underlies all string parsing, canonicalization, and enumeration workflows used in this work. Full implementation details are provided in SI, Text S1.

5.2. String parsing

T-REX strings are parsed with a regex-based state machine that splits each entry into top-level header, ligand, and MAP blocks separated by vertical bars, with nested delimiter handling to accommodate arbitrary SMILES payloads in the ligand list. The MAP block is interpreted as a set of *trans* pairs and singletons over ligand-local "catom" indices, with validation to ensure all referenced ligands and atoms are consistent with the payloads. Full parsing logic is described in SI, Text S2.

5.3. Canonicalization algorithm

Canonicalization proceeds *via* a multi-stage workflow that standardizes ligand payloads with RDKit, remaps coordinating atom indices, ranks ligands by a hierarchical priority tuple (denticity, hapticity, donor atomic number, coordination mass, ligand mass, structural hash), and then lexicographically minimizes the MAP block. Intra- and inter-ligand symmetry, as well as ambiguous geometry flags (*e.g.*, octahedral *vs.* trigonal prismatic in 6-coordinate, 3-pair cases), are resolved to ensure that each chemical species maps to a unique T-REX string. Full algorithmic details, including graph-automorphism handling and WL-hash definitions, are given in SI, Text S3.

Canonicalization added modest overhead during dataset processing. On the 63 375-complex tmQMg-derived dataset, end-to-end parsing and full RDKit-aware canonicalization required a median of 1.56 ms per complex, corresponding to 101 s total runtime, while a lightweight canonicalization without RDKit required a median of 0.077 ms per complex and 5.0 s total runtime. Timing was calculated on a MacBook Air with an M4 processor using a single core.

5.4. 3D structure to T-REX translation

XYZ geometries are converted to non-canonical T-REX strings using a multi-step translation pipeline. First, connectivity is inferred from interatomic distances, the metal center is disconnected from the ligand fragments, and ligand charges and metal oxidation states are assigned *via* an extended Hückel workflow following Rasmussen *et al.*[17] The idealized coordination geometry label is previously assigned from molSimplify's RMSD-based polyhedron matching.[62,63] The geometry label determines the expected number of *trans* pairs, avoiding the need for a fixed angular cutoff. The corresponding *trans* pairs are selected from the largest site-metal-site angles computed from coordination-site centroid vectors, while the remaining sites are assigned as singletons. The resulting ligand identities, coordination sites,[64] hapticity groups, *trans*-pair map, geometry label, and chirality flag are assembled into a valid T-REX string, which is subsequently passed to the canonicalizer described above. A complete description of the translation pipeline, and failure modes is provided in SI, Text S4.

5.5. Chirality detection

Metal-centered chirality is computed in two stages. Point-central chirality (@/@@) is evaluated first: four coordination sites are selected according to geometry-specific rules (tetrahedral, trigonal bipyramidal, square pyramidal, or octahedral), and the sign of their scalar triple product determines handedness. Before computation, achirality is detected through geometry-specific equivalence checks on sites and *trans*-pair sets using the same canonical rank machinery as the canonicalization algorithm. If the complex is not point-chiral, helical chirality (Δ/Λ) is evaluated for octahedral complexes bearing multidentate ligands: tris-bidentate and *cis*-bis-bidentate cases use a propeller method in which the twist of chelate bite vectors around a pseudo-symmetry axis determines handedness, while *fac–fac* bis-tridentate cases measure the angular offset between the two triangular coordination faces. The chirality flag is assigned during 3D-to-T-REX conversion using coordinates from the source structure and is preserved through canonicalization. Full algorithmic details, point-selection rules, and achirality conditions for each geometry are given in SI, Text S5.

5.6. Isomer classification

Isomer relationships between complexes sharing the same metal, oxidation state, and ligand composition are classified through a hierarchical comparison scheme. Each T-REX structure is reduced to a five-level fingerprint: (1) a composition hash encoding the metal, oxidation state, spin, and ligand multiset; (2) a site hash adding the normalized coordinating-atom sets per ligand type; (3) a geometry hash adding the coordination geometry flag; (4) a map hash adding the canonical *trans*-pair arrangement; and (5) a full hash adding the chirality flag.

Structures are grouped by composition hash, then classified by descending through the hierarchy: structures differing at the site level are linkage isomers (same denticity, different binding atoms) or hemilabile isomers (different denticity), those differing at the geometry level are geometric isomers, those differing at the map level are coordination isomers, and those matching through the map but differing in chirality flag are enantiomers. Full algorithmic details are given in SI, Text S6.

5.7. Coordination isomer enumeration

To enumerate coordination isomers, we apply a combinatorial engine that permutes the MAP block for a given T-REX string while preserving multidentate *cis* relationships and forbidding chemically impossible *trans*-chelates, generating all unique assignments of sites into *trans* pairs and singletons consistent with a specified coordination number/geometry. Symmetry-aware deduplication based on ligand identity and internal site symmetry collapses redundant permutations, and each valid map is recombined with the original header and ligand list to yield a set of canonical T-REX strings for all accessible isomers. Algorithmic details, complexity analysis, and validation examples are given in SI, Text S7.

5.8. Ligand substitution and generative expansion

Ligand-centric generative workflows begin from a T-REX dataset, from which we build a ligand registry and classify ligands into interchangeable "strict" and "relaxed" classes based on denticity, hapticity, charge, donor type, and internal *trans*-count constraints. Class-based single and double substitutions are then applied to parent complexes, with canonicalization-based deduplication and optional SMARTS filters enforcing geometry compatibility and domain-specific chemical rules (*e.g.*, *cis*-N motifs). The full indexing scheme, substitution rules, and enumeration protocols are detailed in SI, Text S8.

5.9. Graph and hypergraph neural networks

For ML experiments, T-REX strings are converted to RDKit molecular graphs and used to train five architectures on tmQMg DFT labels (HOMO, LUMO, HOMO–LUMO gap, dipole moment) with a random 80/10/10 split. All models share identical atom featurizations (one-hot element encodings, RDKit-derived properties, Pauling electronegativity, organic and metal-centered chirality tags) and bond featurizations (bond type, conjugation, aromaticity, ring membership, dative flag). The bond-only baseline (MPNN) uses GINE convolutions over this graph. Four geometry-aware architectures augment the bond graph with hyperedges constructed from the T-REX *trans*-pair map, where each hyperedge connects two coordination sites (A and B) through the metal center (M), labeled as *cis* or *trans* with a discrete ideal-angle class derived from the coordination geometry. HyperMPNN uses attention-based GRU pooling with a shared site scorer and routes hyperedge information back through atom features before graph-level pooling. LF-GNN uses the same hyperedge processor but removes metal–ligand dative bonds from the bond graph (forcing all coordination information through hyperedges), adds per-site hyperedges for all metal–ligand one-body terms, and critically maintains a direct hyperedge-to-head readout that concatenates atom-level mean pooling with hyperedge-level mean and max pooling. DeepSets replaces the attention-based processor with a permutation-invariant aggregation (shared site encoder for A/B roles, separate metal encoder, sum pooling) while retaining the direct readout. Virtual Node eliminates dedicated hyperedge processing by encoding each hyperedge as a virtual graph node connected to its member atoms *via* typed edges, with virtual nodes pooled separately at readout. For each architecture and target property, hyperparameters were optimized using Optuna, and final results are reported as ensembles of five independently seeded runs trained with AdamW, cosine annealing, exponential moving average, and mixed-precision training. Detailed architectures, feature definitions, loss functions, and training schedules are provided in SI, Text S9.

Author contributions

I.K. was responsible for conceptualization, methodology, software, investigation, validation, data curation, formal analysis, visualization, resources, supervision, writing – original draft, and review & editing. D.D. contributed to formal analysis, visualization, data curation, and writing – review & editing.

Conflicts of interest

The authors declare no competing financial interest.

Data availability

The code supporting this study is openly available on GitHub at https://github.com/iliak14/trex/. An archived version of the software, datasets, trained models, and workflows, has been deposited in Zenodo with the DOI: https://doi.org/10.5281/zenodo.20046571. Earlier versions of the software, datasets, trained models, and workflows are accessible with the DOI: https://doi.org/10.5281/zenodo.19103065 and https://doi.org/10.5281/zenodo.17905257. The tmQMg, tmCAT, tmPHOTO, tmSCO, and tmBIO datasets are available from their original sources as cited in the manuscript. The T-REX-ent enantiomer dataset that contains enantiomer pair isomers is deposited separately with the DOI: https://doi.org/10.5281/zenodo.19103243. The trex-notation package is also available on PyPI: https://pypi.org/project/trex-notation/0.1.0/.

Supplementary information (SI): an example of a linear T-REX; an example of a bent T-REX; an example of a trigonal planar T-REX; an example of a seesaw T-REX; an example of a square pyramidal T-REX; an example of a trigonal bipyramidal T-REX; an example of a piano-stool complex T-REX; the distribution of oxidation states for the 20 most common metals; distribution of coordination geometries and coordination numbers for the 20 most frequently occurring metals in the tmQMg dataset; examples of linkage isomers; examples of hemilabile isomers; examples of geometry isomers; ligand classification and generative datasets; t-SNE visualization of the chemical space covered by metal hydrides; t-SNE visualization

of the chemical space covered by tmSCO expansion; parity plots of the test set for the HOMO energy across 5 independent seeds and the ensemble average using the MPNN architecture; error distribution of the test set for the HOMO energy across 5 independent seeds with the MPNN architecture; parity plots of the test set for the HOMO energy across 5 independent seeds and the ensemble average using the HyperMPNN architecture; error distribution of the test set for the HOMO energy across 5 independent seeds with the HyperMPNN architecture; parity plots of the test set for the HOMO energy across 5 independent seeds and the ensemble average using the LF-GNN architecture; error distribution of the test set for the HOMO energy across 5 independent seeds with the LF-GNN architecture; parity plots of the test set for the HOMO energy across 5 independent seeds and the ensemble average using the DeepSets LF-GNN architecture; error distribution of the test set for the HOMO energy across 5 independent seeds with the DeepSets LF-GNN architecture; parity plots of the test set for the HOMO energy across 5 independent seeds and the ensemble average using the Virtual Node LF-GNN architecture; error distribution of the test set for the HOMO energy across 5 independent seeds with the Virtual Node LF-GNN architecture; parity plots of the test set for the LUMO energy across 5 independent seeds and the ensemble average using the MPNN architecture; error distribution of the test set for the LUMO energy across 5 independent seeds with the MPNN architecture; parity plots of the test set for the LUMO energy across 5 independent seeds and the ensemble average using the HyperMPNN architecture; error distribution of the test set for the LUMO energy across 5 independent seeds with the HyperMPNN architecture; parity plots of the test set for the LUMO energy across 5 independent seeds and the ensemble average using the LF-GNN architecture; error distribution of the test set for the LUMO energy across 5 independent seeds with the LF-GNN architecture; parity plots of the test set for the LUMO energy across 5 independent seeds and the ensemble average using the DeepSets LF-GNN architecture; error distribution of the test set for the LUMO energy across 5 independent seeds with the DeepSets LF-GNN architecture; parity plots of the test set for the LUMO energy across 5 independent seeds and the ensemble average using the Virtual Node LF-GNN architecture; error distribution of the test set for the LUMO energy across 5 independent seeds with the Virtual Node LF-GNN architecture; parity plots of the test set for the HOMO–LUMO Gap across 5 independent seeds and the ensemble average using the MPNN architecture; error distribution of the test set for the HOMO–LUMO Gap across 5 independent seeds with the MPNN architecture; parity plots of the test set for the HOMO–LUMO Gap across 5 independent seeds and the ensemble average using the HyperMPNN architecture; error distribution of the test set for the HOMO–LUMO Gap across 5 independent seeds with the HyperMPNN architecture; parity plots of the test set for the HOMO–LUMO Gap across 5 independent seeds and the ensemble average using the LF-GNN architecture; error distribution of the test set for the HOMO–LUMO Gap across 5 independent seeds with the LF-GNN architecture; parity plots of the test set for the HOMO–LUMO Gap across 5 independent seeds and the ensemble average using the DeepSets LF-GNN architecture; error distribution of the test set for the HOMO–LUMO Gap across 5 independent seeds with the DeepSets LF-GNN architecture; parity plots of the test set for the HOMO–LUMO Gap across 5 independent seeds and the ensemble average using the Virtual Node LF-GNN architecture; error distribution of the test set for the HOMO–LUMO Gap across 5 independent seeds with the Virtual Node LF-GNN architecture; parity plots of the test set for the Dipole Moment across 5 independent seeds and the ensemble average using the MPNN architecture; error distribution of the test set for the Dipole Moment across 5 independent seeds with the MPNN architecture; parity plots of the test set for the Dipole Moment across 5 independent seeds and the ensemble average using the HyperMPNN architecture; error distribution of the test set for the Dipole Moment across 5 independent seeds with the HyperMPNN architecture; parity plots of the test set for the Dipole Moment across 5 independent seeds and the ensemble average using the LF-GNN architecture; error distribution of the test set for the Dipole Moment across 5 independent seeds with the LF-GNN architecture; parity plots of the test set for the Dipole Moment across 5 independent seeds and the ensemble average using the DeepSets LF-GNN architecture; error distribution of the test set for the Dipole Moment across 5 independent seeds with the DeepSets LF-GNN architecture; parity plots of the test set for the Dipole Moment across 5 independent seeds and the ensemble average using the Virtual Node LF-GNN architecture; error distribution of the test set for the Dipole Moment across 5 independent seeds with the Virtual Node LF-GNN architecture; learning curves for HOMO–LUMO Gap prediction; Learning curves for HOMO Energy prediction; learning curves for LUMO energy prediction; isomer-holdout evaluation of T-REX-derived coordination topology; general software model; string parsing; canonicalization algorithm; 3D structure to T-REX translation; chirality detection; isomer classification; coordination isomer enumeration; ligand substitution and generative expansion; graph neural networks. See DOI: https://doi.org/10.1039/d6dd00129g.

Acknowledgements

The authors acknowledge Baylor University for startup funds used to support this work.

References

1 D. S. Wigh, J. M. Goodman and A. A. Lapkin, A Review of Molecular Representation in the Age of Machine Learning, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.*, 2022, **12**(5), e1603, DOI: 10.1002/wcms.1603.

2 B. A. Mahjour and C. W. Coley, RDCanon: A Python Package for Canonicalizing the Order of Tokens in SMARTS Queries, *J. Chem. Inf. Model.*, 2024, **64**(8), 2948–2954, DOI: 10.1021/acs.jcim.4c00138.

3 A. Gu, K. Goel and C. Ré, Efficiently Modeling Long Sequences with Structured State Spaces, *arXiv*, 2022, preprint, arXiv:2111.00396, DOI: 10.48550/arXiv.2111.00396.

4 J. Li, O. Zhang, K. Sun, Y. Wang, X. Guan, D. Bagni, M. Haghighatlari, F. L. Kearns, C. Parks, R. E. Amaro and T. Head-Gordon, Mining for Potent Inhibitors through Artificial Intelligence and Physics: A Unified Methodology for Ligand Based and Structure Based Drug Design, *J. Chem. Inf. Model.*, 2024, **64**(24), 9082–9097, DOI: 10.1021/acs.jcim.4c00634.

5 R. Gómez-Bombarelli, J. N. Wei, D. Duvenaud, J. M. Hernández-Lobato, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T. D. Hirzel, R. P. Adams and A. Aspuru-Guzik, Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules, *ACS Cent. Sci.*, 2018, **4**(2), 268–276, DOI: 10.1021/acscentsci.7b00572.

6 M. A. Skinnider, R. G. Stacey, D. S. Wishart and L. J. Foster, Chemical Language Models Enable Navigation in Sparsely Populated Chemical Space, *Nat. Mach. Intell.*, 2021, **3**(9), 759–770, DOI: 10.1038/s42256-021-00368-1.

7 D. Weininger, SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules, *J. Chem. Inf. Comput. Sci.*, 1988, **28**(1), 31–36, DOI: 10.1021/ci00057a005.

8 S. Heller, A. McNaught, S. Stein, D. Tchekhovskoi and I. Pletnev, InChI - the Worldwide Chemical Structure Identifier Standard, *J. Cheminf.*, 2013, **5**(1), 7, DOI: 10.1186/1758-2946-5-7.

9 N. O'Boyle and A. Dalke, DeepSMILES: An Adaptation of SMILES for Use in Machine-Learning of Chemical Structures, *ChemRxiv*, 2018, DOI: 10.26434/chemrxiv.7097960.v1.

10 M. Krenn, F. Häse, A. Nigam, P. Friederich and A. Aspuru-Guzik, Self-Referencing Embedded Strings (SELFIES): A 100% Robust Molecular String Representation, *Mach. Learn. Sci. Technol.*, 2020, **1**(4), 045024, DOI: 10.1088/2632-2153/aba947.

11 M. S. Umer, M. Nabeel, U. Athar, I. Lynch, A. Afantitis, S. Ullah and M. M. Fraz, Large Language Models Meet Molecules: A Systematic Review of Advances and Challenges in AI-Driven Cheminformatics, *Arch. Comput. Methods Eng.*, 2025, **33**, 4867–4908, DOI: 10.1007/s11831-025-10437-y.

12 K. D. Vogiatzis, M. V. Polynski, J. K. Kirkland, J. Townsend, A. Hashemi, C. Liu and E. A. Pidko, Computational Approach to Molecular Catalysis by 3d Transition Metals: Challenges and Opportunities, *Chem. Rev.*, 2019, **119**(4), 2453–2523, DOI: 10.1021/acs.chemrev.8b00361.

13 M. Krenn, Q. Ai, S. Barthel, N. Carson, A. Frei, N. C. Frey, P. Friederich, T. Gaudin, A. A. Gayle, K. M. Jablonka, R. F. Lameiro, D. Lemm, A. Lo, S. M. Moosavi, J. M. Nápoles-Duarte, A. Nigam, R. Pollice, K. Rajan, U. Schatzschneider, P. Schwaller, M. Skreta, B. Smit, F. Strieth-Kalthoff, C. Sun, G. Tom, G. F. von Rudorff, A. Wang, A. D. White, A. Young, R. Yu and A. Aspuru-Guzik, SELFIES and the Future of Molecular String Representations, *Patterns*, 2022, **3**(10), 100588, DOI: 10.1016/j.patter.2022.100588.

14 W. Zou, A. M. Monterroza, Y. Yao, S. Cem Millik, M. M. Cencer, N. J. Rebello, H. K. Beech, M. A. Morris, T.-S. Lin, C. S. Castano, J. A. Kalow, S. L. Craig, A. Nelson, J. S. Moore and B. D. Olsen, Extending BigSMILES to Non-Covalent Bonds in Supramolecular Polymer Assemblies, *Chem. Sci.*, 2022, **13**, 12045–12055, DOI: 10.1039/D2SC02257E.

15 T.-S. Lin, C. W. Coley, H. Mochigase, H. K. Beech, W. Wang, Z. Wang, E. Woods, S. L. Craig, J. A. Johnson, J. A. Kalow, K. F. Jensen and B. D. Olsen, BigSMILES: A Structurally-Based Line Notation for Describing Macromolecules, *ACS Cent. Sci.*, 2019, **5**(9), 1523–1531, DOI: 10.1021/acscentsci.9b00476.

16 T.-S. Lin, N. J. Rebello, G.-H. Lee, M. A. Morris and B. D. Olsen, Canonicalizing BigSMILES for Polymers with Defined Backbones, *ACS Polym. Au*, 2022, **2**(6), 486–500, DOI: 10.1021/acspolymersau.2c00009.

17 M. H. Rasmussen, M. Strandgaard, J. Seumer, L. K. Hemmingsen, A. Frei, D. Balcells and J. H. Jensen, SMILES All around: Structure to SMILES Conversion for Transition Metal Complexes, *J. Cheminf.*, 2025, **17**(1), 63, DOI: 10.1186/s13321-025-01008-1.

18 K. Lee, S. Park, M. Park and W. Y. Kim, MetalloGen: Automated 3D Conformer Generation for Diverse Coordination Complexes, *J. Chem. Inf. Model.*, 2025, **65**(21), 11878–11891, DOI: 10.1021/acs.jcim.5c02074.

19 J. P. Janet and H. J. Kulik, Resolving Transition Metal Chemical Space: Feature Selection for Machine Learning and Structure–Property Relationships, *J. Phys. Chem. A*, 2017, **121**(46), 8939–8954, DOI: 10.1021/acs.jpca.7b08750.

20 J. W. Toney, R. G. St. Michel, A. G. Garrison, I. Kevlishvili and H. J. Kulik, Graph Neural Networks for Predicting Metal–Ligand Coordination of Transition Metal Complexes, *Proc. Natl. Acad. Sci. U. S. A.*, 2025, **122**(41), e2415658122, DOI: 10.1073/pnas.2415658122.

21 J. P. Janet, C. Duan, A. Nandy, F. Liu and H. J. Kulik, Navigating Transition-Metal Chemical Space: Artificial Intelligence for First-Principles Design, *Acc. Chem. Res.*, 2021, **54**(3), 532–545, DOI: 10.1021/acs.accounts.0c00686.

22 I. Kevlishvili, J. Vakil, D. W. Kastner, X. Huang, S. L. Craig and H. J. Kulik, High-Throughput Discovery of Ferrocene Mechanophores with Enhanced Reactivity and Network Toughening, *ACS Cent. Sci.*, 2025, **11**(10), 1839–1851, DOI: 10.1021/acscentsci.5c00707.

23 J. Paul Janet, C. Duan, T. Yang, A. Nandy and H. A. J. Kulik, Quantitative Uncertainty Metric Controls Error in Neural Network-Driven Chemical Discovery, *Chem. Sci.*, 2019, **10**(34), 7913–7922, DOI: 10.1039/C9SC02298H.

24 I. Kevlishvili, C. Duan and H. J. Kulik, Classification of Hemilabile Ligands Using Machine Learning, *J. Phys. Chem. Lett.*, 2023, **14**(49), 11100–11109, DOI: 10.1021/acs.jpclett.3c02828.

25 M. G. Taylor, A. Nandy, C. C. Lu and H. J. Kulik, Deciphering Cryptic Behavior in Bimetallic Transition-Metal Complexes with Machine Learning, *J. Phys. Chem. Lett.*, 2021, **12**(40), 9812–9820, DOI: 10.1021/acs.jpclett.1c02852.

26 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, A

Comprehensive Discovery Platform for Organophosphorus Ligands for Catalysis, *J. Am. Chem. Soc.*, 2022, **144**(3), 1205–1217, DOI: 10.1021/jacs.1c09718.

27 S.-S. Chen, Z. Meyer, B. Jensen, A. Kraus, A. Lambert and D. H. Ess, ReaLigands: A Ligand Library Cultivated from Experiment and Intended for Molecular Computational Catalyst Design, *J. Chem. Inf. Model.*, 2023, **63**(23), 7412–7422, DOI: 10.1021/acs.jcim.3c01310.

28 J. Jover, N. Fey, J. N. Harvey, G. C. Lloyd-Jones, A. G. Orpen, G. J. J. Owen-Smith, P. Murray, D. R. J. Hose, R. Osborne and M. Purdie, Expansion of the Ligand Knowledge Base for Monodentate P-Donor Ligands (LKB-P), *Organometallics*, 2010, **29**(23), 6245–6258, DOI: 10.1021/om100648v.

29 R. A. Mansson, A. H. Welsh, N. Fey and A. G. Orpen, Statistical Modeling of a Ligand Knowledge Base, *J. Chem. Inf. Model.*, 2006, **46**(6), 2591–2600, DOI: 10.1021/ci600212t.

30 J. Jover, N. Fey, J. N. Harvey, G. C. Lloyd-Jones, A. G. Orpen, G. J. J. Owen-Smith, P. Murray, D. R. J. Hose, R. Osborne and M. Purdie, Expansion of the Ligand Knowledge Base for Chelating P,P-Donor Ligands (LKB-PP), *Organometallics*, 2012, **31**(15), 5302–5306, DOI: 10.1021/om300312t.

31 A. G. Garrison, J. Heras-Domingo, J. R. Kitchin, G. dos Passos Gomes, Z. W. Ulissi and S. M. Blau, Applying Large Graph Neural Networks to Predict Transition Metal Complex Energies Using the tmQM_wB97MV Data Set, *J. Chem. Inf. Model.*, 2023, **63**(24), 7642–7654, DOI: 10.1021/acs.jcim.3c01226.

32 M. Jones, G, A. Smith, B., K. Kirkland, J. and D. Vogiatzis, K. Data-Driven Ligand Field Exploration of Fe(Iv)–Oxo Sites for C–H Activation, *Inorg. Chem. Front.*, 2023, **10**(4), 1062–1075, DOI: 10.1039/D2QI01961B.

33 H. Kneiding, R. Lukin, L. Lang, S. Reine, T. Bondo Pedersen, R. D. Bin and D. Balcells, Deep Learning Metal Complex Properties with Natural Quantum Graphs, *Digital Discovery*, 2023, **2**, 618–633.

34 J. Lu, Z. Song, Q. Zhao, Y. Du, Y. Cao, H. Jia and C. Duan, Generative Design of Functional Metal Complexes Utilizing the Internal Knowledge and Reasoning Capability of Large Language Models, *J. Am. Chem. Soc.*, 2025, **147**(36), 32377–32388, DOI: 10.1021/jacs.5c02097.

35 C. Bilodeau, W. Jin, T. Jaakkola, R. Barzilay and K. F. Jensen, Generative Models for Molecular Discovery: Recent Advances and Challenges, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.*, 2022, **12**(5), e1608, DOI: 10.1002/wcms.1608.

36 M. Strandgaard, T. Linjordet, H. Kneiding, A. L. Burnage, A. Nova, J. H. Jensen and D. Balcells, A Deep Generative Model for the Inverse Design of Transition Metal Ligands and Complexes, *JACS Au*, 2025, **5**(5), 2294–2308, DOI: 10.1021/jacsau.5c00242.

37 O. Schilter, A. Vaucher, P. Schwaller and T. Laino, Designing Catalysts with Deep Generative Models and Computational Data. A Case Study for Suzuki Cross Coupling Reactions, *Digital Discovery*, 2023, **2**(3), 728–735, DOI: 10.1039/D2DD00125J.

38 Y. Liu, J. Cavanagh, K. Sun, J. Toney, C.-Y. Yuan, A. Smith, R. S. M. Ii, P. Graggs, F. D. Toste, H. Kulik and T. Head-Gordon, Exploring Transition Metal Complexes with Large Language Models, *ChemRxiv*, 2025, DOI: 10.26434/chemrxiv-2025-hm3zb.

39 M. Bran, A., S. Cox, O. Schilter, C. Baldassari, A. D. White and P. Schwaller, Augmenting Large Language Models with Chemistry Tools, *Nat. Mach. Intell.*, 2024, **6**(5), 525–535, DOI: 10.1038/s42256-024-00832-8.

40 Y. Kang and J. Kim, ChatMOF: An Artificial Intelligence System for Predicting and Generating Metal-Organic Frameworks Using Large Language Models, *Nat. Commun.*, 2024, **15**(1), 4705, DOI: 10.1038/s41467-024-48998-4.

41 D. A. Boiko, R. MacKnight, B. Kline and G. Gomes, Autonomous Chemical Research with Large Language Models, *Nature*, 2023, **624**(7992), 570–578, DOI: 10.1038/s41586-023-06792-0.

42 Y. Wang, H. Zhao, S. Sciabola and W. Wang, cMolGPT: A Conditional Generative Pre-Trained Transformer for Target-Specific De Novo Molecular Generation, *Molecules*, 2023, **28**(11), 4430, DOI: 10.3390/molecules28114430.

43 Z. Zheng, O. Zhang, C. Borgs, J. T. Chayes and O. M. Yaghi, ChatGPT Chemistry Assistant for Text Mining and the Prediction of MOF Synthesis, *J. Am. Chem. Soc.*, 2023, **145**(32), 18048–18062, DOI: 10.1021/jacs.3c05819.

44 Z. Zheng, AI and Chemistry in Action: Transforming Crystallization for Scalable Water Harvesting Solutions, *ACS Cent. Sci.*, 2024, **10**(12), 2173–2174, DOI: 10.1021/acscentsci.4c01838.

45 I. Kevlishvili, R. G. S. G. Michel, A. Garrison, J. W. Toney, H. Adamji, H. Jia, Y. Román-Leshkov and J. Kulik, Leveraging Natural Language Processing to Curate the tmCAT, tmPHOTO, tmBIO, and tmSCO Datasets of Functional Transition Metal Complexes, *Faraday Discuss.*, 2025, **256**, 275–303, DOI: 10.1039/D4FD00087K.

46 I. Kevlishvili, C. Duan and H. J. Kulik, Classification of Hemilabile Ligands Using Machine Learning, *J. Phys. Chem. Lett.*, 2023, **14**(49), 11100–11109, DOI: 10.1021/acs.jpclett.3c02828.

47 J. W. Toney, R. G. St. Michel, A. G. Garrison, I. Kevlishvili and H. J. Kulik, Identifying Dynamic Metal–Ligand Coordination Modes with Ensemble Learning, *J. Am. Chem. Soc.*, 2025, **147**(52), 48218–48234, DOI: 10.1021/jacs.5c17169.

48 G. Moldagulov, K. Lee, S. Nurgaliyev, A. Salem, A. Kuznietsov and B. A. Grzybowski, Hybrid Computational Strategy for Predicting Complex Ligand–Metal Architectures, *Angew. Chem., Int. Ed.*, 2026, **65**, e24655, DOI: 10.1002/anie.202524655.

49 I. Kevlishvili, *T-REX-Ent: Enantiomer Pairs of Chiral TMCs*, 2026, DOI: 10.5281/zenodo.19103243.

50 Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang and P. S. Yu, A Comprehensive Survey on Graph Neural Networks, *IEEE Transact. Neural Networks Learn. Syst.*, 2021, **32**(1), 4–24, DOI: 10.1109/TNNLS.2020.2978386.

51 R. Brossard, O. Frigo and D. Dehaene, Graph Convolutions That Can Finally Model Local Structure, *arXiv*, 2021, preprint, arXiv:2011.15069, DOI: 10.48550/arXiv.2011.15069.

52 W. Du, S. Zhang, Z. Cai, X. Li, Z. Liu, J. Fang, J. Wang, X. Wang and Y. Wang, Molecular Merged Hypergraph Neural Network for Explainable Solvation Gibbs Free

Energy Prediction, *Research*, 2025, **8**, 0740, DOI: 10.34133/research.0740.

53 Y. Feng, H. You, Z. Zhang, R. Ji and Y. Gao, Hypergraph Neural Networks, *Proc. AAAI Conf. Artif. Intell.*, 2019, **33**(01), 3558–3565, DOI: 10.1609/aaai.v33i01.33013558.

54 Y. Lin, Q. Yuan, Q. Qiu and S. Lian, MolHyper: Hypergraph-Enhanced Graph Network for Accurate Molecular Property Prediction. in *2024 14th International Conference on Information Technology in Medicine and Education (ITME)*, 2024, pp 659–663, DOI: 10.1109/ITME63426.2024.00135.

55 J. Chen and P. Schwaller, Molecular Hypergraph Neural Networks, *J. Chem. Phys.*, 2024, **160**(14), 144307, DOI: 10.1063/5.0193557.

56 K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk and Y. Bengio, Learning Phrase Representations Using RNN Encoder–Decoder for Statistical Machine Translation. in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, ed A., Moschitti, B., Pang, W., Daelemans, Association for Computational Linguistics, Doha, Qatar, 2014, pp 1724–1734, DOI: 10.3115/v1/D14-1179.

57 M. Zaheer, S. Kottur, S. Ravanbakhsh, B. Poczos, R. R. Salakhutdinov and A. J. Smola, Deep Sets. in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2017, Vol. 30.

58 J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals and G. E. Dahl Neural Message Passing for Quantum Chemistry. in *Proceedings of the 34th International Conference on Machine Learning*, PMLR, 2017, pp 1263–1272.

59 P. van Gerwen, K. R. Briling, C. Bunne, V. R. Somnath, R. Laplaza, A. Krause and C. Corminboeuf, 3DReact: Geometric Deep Learning for Chemical Reactions, *J. Chem. Inf. Model.*, 2024, **64**(15), 5771–5785, DOI: 10.1021/acs.jcim.4c00104.

60 K. Atz, F. Grisoni and G. Schneider, Geometric Deep Learning on Molecular Representations, *Nat. Mach. Intell.*, 2021, **3**(12), 1023–1032, DOI: 10.1038/s42256-021-00418-8.

61 S. Batzner, A. Musaelian, L. Sun, M. Geiger, J. P. Mailoa, M. Kornbluth, N. Molinari, T. E. Smidt and B. Kozinsky, E(3)-Equivariant Graph Neural Networks for Data-Efficient and Accurate Interatomic Potentials, *Nat. Commun.*, 2022, **13**(1), 2453, DOI: 10.1038/s41467-022-29939-5.

62 E. I. Ioannidis, T. Z. H. Gani and H. J. Kulik, molSimplify: A Toolkit for Automating Discovery in Inorganic Chemistry, *J. Comput. Chem.*, 2016, **37**(22), 2106–2117, DOI: 10.1002/jcc.24437.

63 G. Terrones, R. S. Michel, J. Toney, A. Ball, Y. Wang, A. Garrison, A. Nandy, R. Meyer, F. Edholm, C. Oh, S. Pujet, D. Chu, D. Muhammetgulyyev and H. Kulik, molSimplify 2.0: Improved Structure Generation for Automating Discovery in Inorganic Molecular and Reticular Chemistry, *J. Chem. Inf. Model.*, 2026, **66**(5), 2753–2767, DOI: 10.26434/chemrxiv-2025-h8gff.

64 I. Kevlishvili and D. Dorabawila, *Data for Taming T-REX: A Canonical Language for Geometry-Aware Generative Design of Transition-Metal Complexes*, 2026, DOI: 10.5281/zenodo.20046571.