23 March 2026

Taming T-REX: A Canonical Language for Geometry-Aware Generative Design of Transition-Metal Complexes

Ilia Kevlishvili[1], Devmin Dorabawila

1. Baylor University

Abstract

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.

Taming T-REX: A Canonical Language for Geometry-Aware Generative Design of Transition-Metal Complexes

Ilia Kevlishvili[1,*], Devmin Dorabawila[1]

[1]Department of Chemistry and Biochemistry, Baylor University, Waco, Texas, USA

[*]Corresponding author. email: ilia_kevlishvili@baylor.edu

Abstract

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.

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 species maps to exactly one deterministic string, ensuring database integrity and preventing duplicate bias in machine learning models.

In this work, we (i) introduce Trans-pair Relations EXpression (T-REX), a canonical line notation for monometallic CN≤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 (Figure 1, Supporting Information Figures S1-S7).

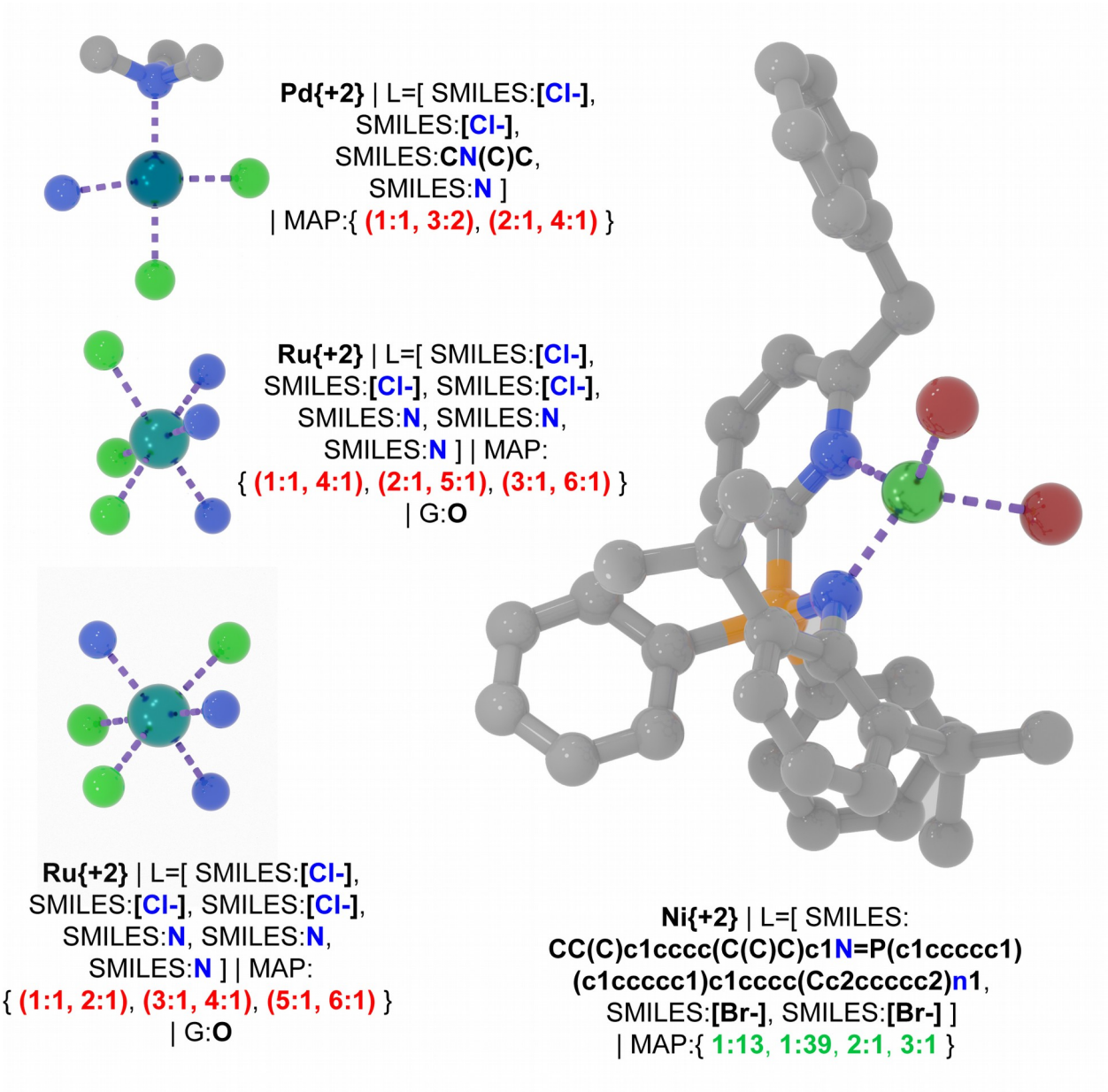


Figure 1. Examples of T-REX strings for square planar (top left), tetrahedral (right), *fac*-octahedral, and *mer*-octahedral complexes. Coordinating atoms are highlighted in blue, *trans*-pairs are highlighted in red, "singles" are highlighted in green.

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=[lig`$_1$`, lig`$_2$`, … ]`. 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 majority of coordination geometries and all *cis* relationships up to CN ≤ 6 (Figure 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 Supporting Information, 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 (Figure 2).

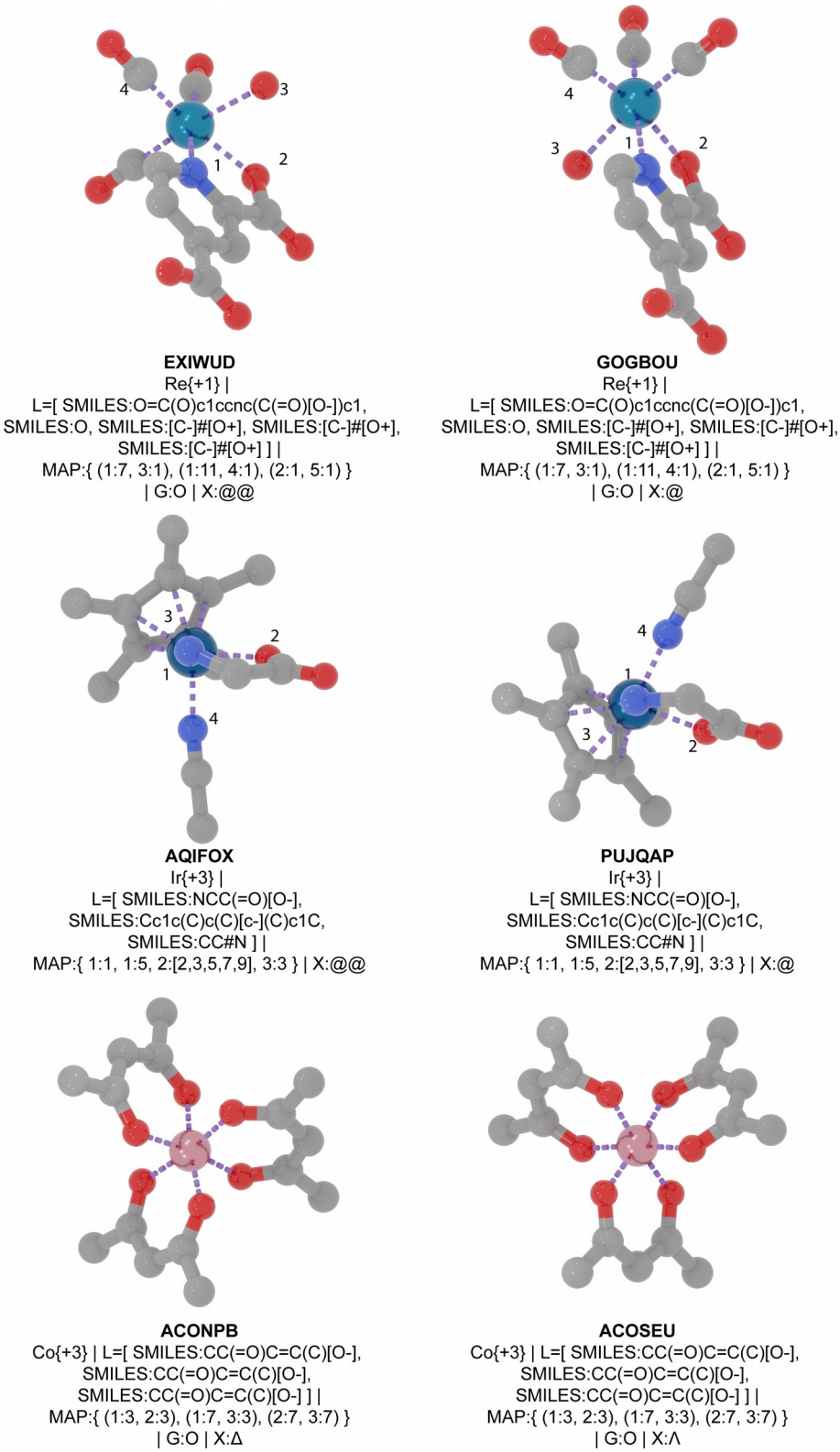


Figure 2. Examples of chiral molecules represented in T-REX. Atoms used to compute point-central chirality are labeled.

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 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. Future work will focus on addressing these shortcomings and expanding grammar for the multinuclear MUL-T-REX, which generalizes the same block structure while preserving the canonical, edit-friendly design.

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

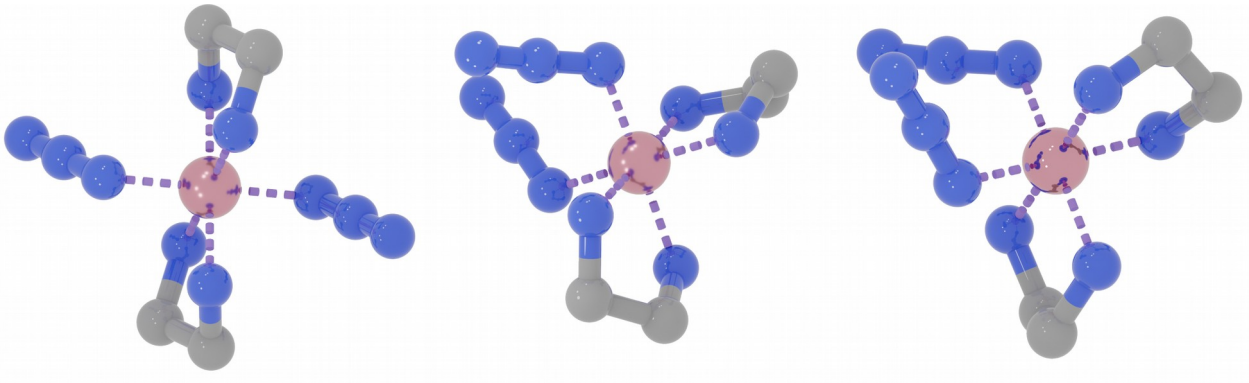
We validated the representation's coverage by converting the tmQMg dataset[33], which is a comprehensive library of DFT-optimized transition-metal complex geometries, into canonical T-REX strings. Of the 74,547 starting structures, the conversion pipeline successfully parsed 72,733 (97.5%), with rejections 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. 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 2,007 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[45] (tmCAT, tmPHOTO, tmBIO, and tmSCO), yielding 18,855, 4,061, 2,542, and 8,209 unique complexes, respectively.

The curated tmQMg library spans 52 transition metals across oxidation states from −3 to +7 (Supporting Information Figure 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 (Supporting Information Figure 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 Figures 2 and 3, respectively, and other examples are shown in Supporting Information Figures S10-S12.



ACINUX
Co{+3} | L=[SMILES:NCCN, SMILES:NCCN, SMILES:N#[N+][N-2], SMILES:N#[N+][N-2]] | MAP:{ (1:1, 2:1), (1:4, 2:4), (3:3, 4:3) } | G:O
N#[N+][N-2]->[Co+3]12(<-[N-2][N+]#N)(<-NCCN->1)<-NCCN->2

LIKMOL
Co{+3} | L=[SMILES:NCCN, SMILES:NCCN, SMILES:N#[N+][N-2], SMILES:N#[N+][N-2]] | MAP:{ (1:1, 2:1), (1:4, 3:3), (2:4, 4:3) } | G:O | X:Δ
N#[N+][N-2]->[Co+3]12(<-[N-2][N+]#N)(<-NCCN->1)<-NCCN->2

YASLIO
Co{+3} | L=[SMILES:NCCN, SMILES:NCCN, SMILES:N#[N+][N-2], SMILES:N#[N+][N-2]] | MAP:{ (1:1, 2:1), (1:4, 3:3), (2:4, 4:3) } | G:O | X:Λ
N#[N+][N-2]->[Co+3]12(<-[N-2][N+]#N)(<-NCCN->1)<-NCCN->2

Figure 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.

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., $Co(III)(en)_2(N_3)_2$, Figure 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 (Supporting Information 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 structures. 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 (Supporting Information Figure S13).

Furthermore, the integration of SMARTS logic allows for the imposition of specific chemical rules during generation, as 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 (Figure 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. (Supporting Information Figure S14).

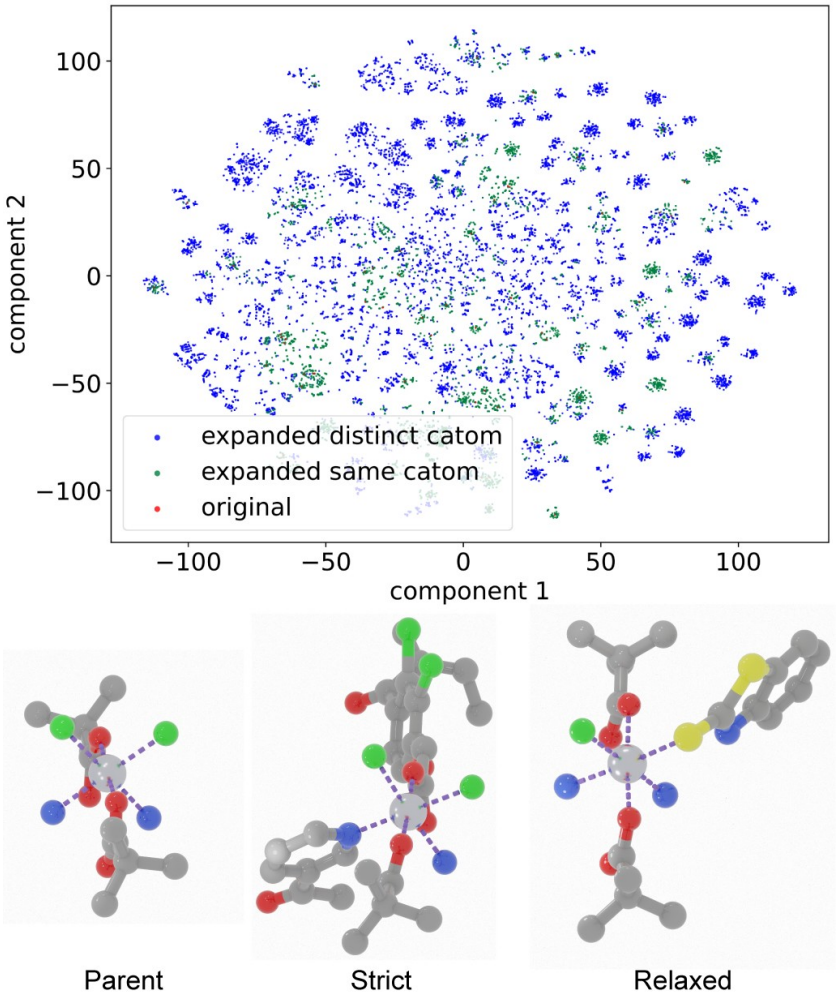


Figure 4. t-SNE visualization of the chemical space covered by the generated *cis*-diamino complexes in tmBIO dataset (top), and Pt (IV) structures from these datasets (bottom). Original parent complexes are shown in red, strict expansion is shown in green, and relaxed expansion is shown in blue. The 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.

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.

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 6,337 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 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 (Supporting Information, 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 (Supporting Information Figures 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 (Supporting Information Figures 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, Supporting Information Figures 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) (Figure 5, Table 1, Supporting Information Figures S45–S54). 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.

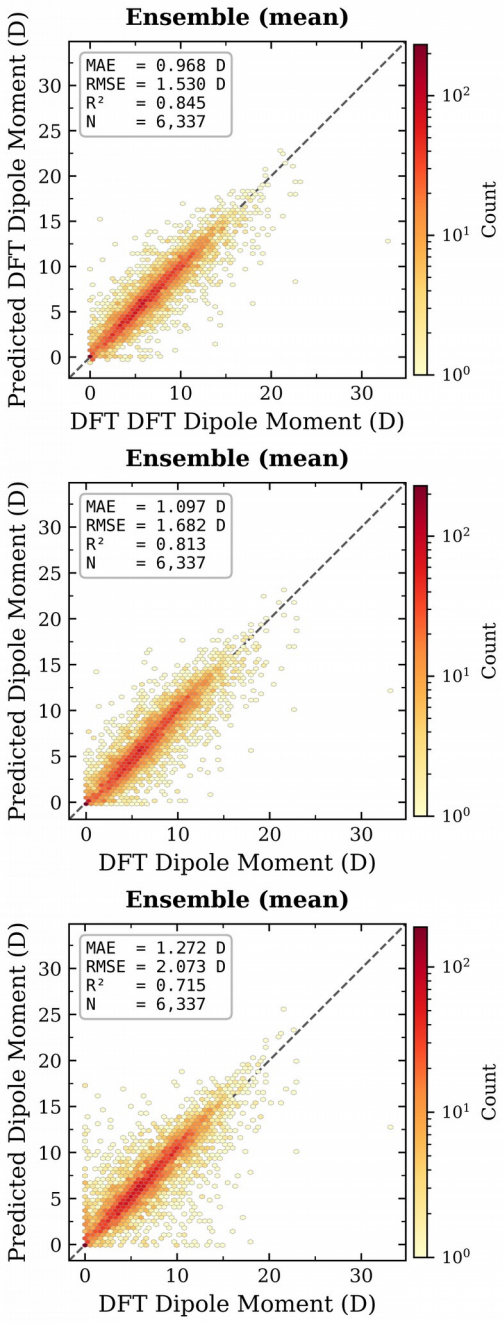


Figure 5. Parity plots of predicted vs calculated dipole moment on the set aside test set using the LF-GNN (top), HyperMPNN (middle), and MPNN (bottom) architectures.

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].

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 (Figure 6, Supporting Information Figures 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.

Figure 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.

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. 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≤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

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 Supporting Information, Text S1.

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 Supporting Information, Text S2.

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 Supporting Information, Text S3.

3D Structure to T-REX Translation

XYZ geometries are converted to non-canonical T-REX strings using a multi-step pipeline that infers connectivity from distances, disconnects the metal and ligand fragments, assigns ligand charges and metal oxidation states via an extended Hückel workflow following Rasmussen et al.[17], and classifies coordination geometry using molSimplify’s RMSD-based polyhedron matching[62,63]. Coordination sites, hapticity groups, and trans pairs are then extracted from metal–ligand vectors and converted 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 Supporting Information, Text S4.

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 Supporting Information, Text S5.

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 Supporting Information, Text S6.

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 Supporting Information, Text S7.

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 Supporting Information, Text S8.

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 Supporting Information, Text S9.

Supporting Information

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; 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.

Corresponding Author

*email: ilia_kevlishvili@baylor.edu phone: 254-710-4272

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENT

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

Data Availability

The T-REX datasets (tmQMg translation, enumerated isomers, T-REX-ent) and model training data are deposited on Zenodo[64]. The T-REX-ent enantiomer dataset is deposited separately[49]. The trex-notation package is available on PyPI. All code is available at https://github.com/iliak14/trex/. References cited in the Supporting. All data are available in the main text, Supporting Information, Zenodo repository, and Github repository: https://github.com/iliak14/trex/ .

References

(1) Wigh, D. S.; Goodman, J. M.; Lapkin, A. A. A Review of Molecular Representation in the Age of Machine Learning. *WIREs Comput. Mol. Sci.* **2022**, *12* (5), e1603. https://doi.org/10.1002/wcms.1603.

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

(3) Gu, A.; Goel, K.; Ré, C. Efficiently Modeling Long Sequences with Structured State Spaces. arXiv August 5, 2022. https://doi.org/10.48550/arXiv.2111.00396.

(4) Li, J.; Zhang, O.; Sun, K.; Wang, Y.; Guan, X.; Bagni, D.; Haghighatlari, M.; Kearns, F. L.; Parks, C.; Amaro, R. E.; Head-Gordon, T. 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. https://doi.org/10.1021/acs.jcim.4c00634.

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

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

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

(8) Heller, S.; McNaught, A.; Stein, S.; Tchekhovskoi, D.; Pletnev, I. InChI - the Worldwide Chemical Structure Identifier Standard. *J. Cheminformatics* **2013**, *5* (1), 7. https://doi.org/10.1186/1758-2946-5-7.
(9) O'Boyle, N.; Dalke, A. DeepSMILES: An Adaptation of SMILES for Use in Machine-Learning of Chemical Structures. ChemRxiv September 19, 2018. https://doi.org/10.26434/chemrxiv.7097960.v1.
(10) Krenn, M.; Häse, F.; Nigam, A.; Friederich, P.; Aspuru-Guzik, A. Self-Referencing Embedded Strings (SELFIES): A 100% Robust Molecular String Representation. *Mach. Learn. Sci. Technol.* **2020**, *1* (4), 045024. https://doi.org/10.1088/2632-2153/aba947.
(11) Umer, M. S.; Nabeel, M.; Athar, U.; Lynch, I.; Afantitis, A.; Ullah, S.; Fraz, M. M. Large Language Models Meet Molecules: A Systematic Review of Advances and Challenges in AI-Driven Cheminformatics. *Arch. Comput. Methods Eng.* **2025**. https://doi.org/10.1007/s11831-025-10437-y.
(12) Vogiatzis, K. D.; Polynski, M. V.; Kirkland, J. K.; Townsend, J.; Hashemi, A.; Liu, C.; Pidko, E. A. Computational Approach to Molecular Catalysis by 3d Transition Metals: Challenges and Opportunities. *Chem. Rev.* **2019**, *119* (4), 2453–2523. https://doi.org/10.1021/acs.chemrev.8b00361.
(13) Krenn, M.; Ai, Q.; Barthel, S.; Carson, N.; Frei, A.; Frey, N. C.; Friederich, P.; Gaudin, T.; Gayle, A. A.; Jablonka, K. M.; Lameiro, R. F.; Lemm, D.; Lo, A.; Moosavi, S. M.; Nápoles-Duarte, J. M.; Nigam, A.; Pollice, R.; Rajan, K.; Schatzschneider, U.; Schwaller, P.; Skreta, M.; Smit, B.; Strieth-Kalthoff, F.; Sun, C.; Tom, G.; Rudorff, G. F. von; Wang, A.; White, A. D.; Young, A.; Yu, R.; Aspuru-Guzik, A. SELFIES and the Future of Molecular String Representations. *Patterns* **2022**, *3* (10). https://doi.org/10.1016/j.patter.2022.100588.
(14) Zou, W.; Monterroza, A. M.; Yao, Y.; Cem Millik, S.; M. Cencer, M.; J. Rebello, N.; K. Beech, H.; A. Morris, M.; Lin, T.-S.; S. Castano, C.; A. Kalow, J.; L. Craig, S.; Nelson, A.; S. Moore, J.; D. Olsen, B. Extending BigSMILES to Non-Covalent Bonds in Supramolecular Polymer Assemblies. **2022**. https://doi.org/10.1039/D2SC02257E.
(15) Lin, T.-S.; Coley, C. W.; Mochigase, H.; Beech, H. K.; Wang, W.; Wang, Z.; Woods, E.; Craig, S. L.; Johnson, J. A.; Kalow, J. A.; Jensen, K. F.; Olsen, B. D. BigSMILES: A Structurally-Based Line Notation for Describing Macromolecules. *ACS Cent. Sci.* **2019**, *5* (9), 1523–1531. https://doi.org/10.1021/acscentsci.9b00476.
(16) Lin, T.-S.; Rebello, N. J.; Lee, G.-H.; Morris, M. A.; Olsen, B. D. Canonicalizing BigSMILES for Polymers with Defined Backbones. *ACS Polym. Au* **2022**, *2* (6), 486–500. https://doi.org/10.1021/acspolymersau.2c00009.
(17) Rasmussen, M. H.; Strandgaard, M.; Seumer, J.; Hemmingsen, L. K.; Frei, A.; Balcells, D.; Jensen, J. H. SMILES All around: Structure to SMILES Conversion for Transition Metal Complexes. *J. Cheminformatics* **2025**, *17* (1), 63. https://doi.org/10.1186/s13321-025-01008-1.
(18) Lee, K.; Park, S.; Park, M.; Kim, W. Y. MetalloGen: Automated 3D Conformer Generation for Diverse Coordination Complexes. *J. Chem. Inf.*

Model. **2025**, *65* (21), 11878–11891. https://doi.org/10.1021/acs.jcim.5c02074.
(19) Janet, J. P.; Kulik, H. J. Resolving Transition Metal Chemical Space: Feature Selection for Machine Learning and Structure–Property Relationships. *J. Phys. Chem. A* **2017**, *121* (46), 8939–8954. https://doi.org/10.1021/acs.jpca.7b08750.
(20) Toney, J. W.; St. Michel, R. G.; Garrison, A. G.; Kevlishvili, I.; Kulik, H. J. Graph Neural Networks for Predicting Metal–Ligand Coordination of Transition Metal Complexes. *Proc. Natl. Acad. Sci.* **2025**, *122* (41), e2415658122. https://doi.org/10.1073/pnas.2415658122.
(21) Janet, J. P.; Duan, C.; Nandy, A.; Liu, F.; Kulik, H. J. Navigating Transition-Metal Chemical Space: Artificial Intelligence for First-Principles Design. *Acc. Chem. Res.* **2021**, *54* (3), 532–545. https://doi.org/10.1021/acs.accounts.0c00686.
(22) Kevlishvili, I.; Vakil, J.; Kastner, D. W.; Huang, X.; Craig, S. L.; Kulik, H. J. High-Throughput Discovery of Ferrocene Mechanophores with Enhanced Reactivity and Network Toughening. *ACS Cent. Sci.* **2025**, *11* (10), 1839–1851. https://doi.org/10.1021/acscentsci.5c00707.
(23) Paul Janet, J.; Duan, C.; Yang, T.; Nandy, A.; J. Kulik, H. A Quantitative Uncertainty Metric Controls Error in Neural Network-Driven Chemical Discovery. *Chem. Sci.* **2019**, *10* (34), 7913–7922. https://doi.org/10.1039/C9SC02298H.
(24) Kevlishvili, I.; Duan, C.; Kulik, H. J. Classification of Hemilabile Ligands Using Machine Learning. *J. Phys. Chem. Lett.* **2023**, *14* (49), 11100–11109. https://doi.org/10.1021/acs.jpclett.3c02828.
(25) Taylor, M. G.; Nandy, A.; Lu, C. C.; Kulik, H. J. Deciphering Cryptic Behavior in Bimetallic Transition-Metal Complexes with Machine Learning. *J. Phys. Chem. Lett.* **2021**, *12* (40), 9812–9820. https://doi.org/10.1021/acs.jpclett.1c02852.
(26) Gensch, T.; dos Passos Gomes, G.; Friederich, P.; Peters, E.; Gaudin, T.; Pollice, R.; Jorner, K.; Nigam, A.; Lindner-D'Addario, M.; Sigman, M. S.; Aspuru-Guzik, A. A Comprehensive Discovery Platform for Organophosphorus Ligands for Catalysis. *J. Am. Chem. Soc.* **2022**, *144* (3), 1205–1217. https://doi.org/10.1021/jacs.1c09718.
(27) Chen, S.-S.; Meyer, Z.; Jensen, B.; Kraus, A.; Lambert, A.; Ess, D. H. ReaLigands: A Ligand Library Cultivated from Experiment and Intended for Molecular Computational Catalyst Design. *J. Chem. Inf. Model.* **2023**, *63* (23), 7412–7422. https://doi.org/10.1021/acs.jcim.3c01310.
(28) Jover, J.; Fey, N.; Harvey, J. N.; Lloyd-Jones, G. C.; Orpen, A. G.; Owen-Smith, G. J. J.; Murray, P.; Hose, D. R. J.; Osborne, R.; Purdie, M. Expansion of the Ligand Knowledge Base for Monodentate P-Donor Ligands (LKB-P). *Organometallics* **2010**, *29* (23), 6245–6258. https://doi.org/10.1021/om100648v.
(29) Mansson, R. A.; Welsh, A. H.; Fey, N.; Orpen, A. G. Statistical Modeling of a Ligand Knowledge Base. *J. Chem. Inf. Model.* **2006**, *46* (6), 2591–2600. https://doi.org/10.1021/ci600212t.

(30) Jover, J.; Fey, N.; Harvey, J. N.; Lloyd-Jones, G. C.; Orpen, A. G.; Owen-Smith, G. J. J.; Murray, P.; Hose, D. R. J.; Osborne, R.; Purdie, M. Expansion of the Ligand Knowledge Base for Chelating P,P-Donor Ligands (LKB-PP). *Organometallics* **2012**, *31* (15), 5302–5306. https://doi.org/10.1021/om300312t.
(31) Garrison, A. G.; Heras-Domingo, J.; Kitchin, J. R.; dos Passos Gomes, G.; Ulissi, Z. W.; Blau, S. M. 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. https://doi.org/10.1021/acs.jcim.3c01226.
(32) M. Jones, G.; A. Smith, B.; K. Kirkland, J.; 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. https://doi.org/10.1039/D2QI01961B.
(33) Kneiding, H.; Lukin, R.; Lang, L.; Reine, S.; Bondo Pedersen, T.; Bin, R. D.; Balcells, D. Deep Learning Metal Complex Properties with Natural Quantum Graphs. **2023**. https://doi.org/10.1039/D2DD00129B.
(34) Lu, J.; Song, Z.; Zhao, Q.; Du, Y.; Cao, Y.; Jia, H.; Duan, C. 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. https://doi.org/10.1021/jacs.5c02097.
(35) Bilodeau, C.; Jin, W.; Jaakkola, T.; Barzilay, R.; Jensen, K. F. Generative Models for Molecular Discovery: Recent Advances and Challenges. *WIREs Comput. Mol. Sci.* **2022**, *12* (5), e1608. https://doi.org/10.1002/wcms.1608.
(36) Strandgaard, M.; Linjordet, T.; Kneiding, H.; Burnage, A. L.; Nova, A.; Jensen, J. H.; Balcells, D. A Deep Generative Model for the Inverse Design of Transition Metal Ligands and Complexes. *JACS Au* **2025**, *5* (5), 2294–2308. https://doi.org/10.1021/jacsau.5c00242.
(37) Schilter, O.; Vaucher, A.; Schwaller, P.; Laino, T. Designing Catalysts with Deep Generative Models and Computational Data. A Case Study for Suzuki Cross Coupling Reactions. *Digit. Discov.* **2023**, *2* (3), 728–735. https://doi.org/10.1039/D2DD00125J.
(38) Liu, Y.; Cavanagh, J.; Sun, K.; Toney, J.; Yuan, C.-Y.; Smith, A.; Ii, R. S. M.; Graggs, P.; Toste, F. D.; Kulik, H.; Head-Gordon, T. Exploring Transition Metal Complexes with Large Language Models. ChemRxiv November 17, 2025. https://doi.org/10.26434/chemrxiv-2025-hm3zb.
(39) M. Bran, A.; Cox, S.; Schilter, O.; Baldassari, C.; White, A. D.; Schwaller, P. Augmenting Large Language Models with Chemistry Tools. *Nat. Mach. Intell.* **2024**, *6* (5), 525–535. https://doi.org/10.1038/s42256-024-00832-8.
(40) Kang, Y.; Kim, J. ChatMOF: An Artificial Intelligence System for Predicting and Generating Metal-Organic Frameworks Using Large Language Models. *Nat. Commun.* **2024**, *15* (1), 4705. https://doi.org/10.1038/s41467-024-48998-4.

(41) Boiko, D. A.; MacKnight, R.; Kline, B.; Gomes, G. Autonomous Chemical Research with Large Language Models. *Nature* **2023**, *624* (7992), 570–578. https://doi.org/10.1038/s41586-023-06792-0.
(42) Wang, Y.; Zhao, H.; Sciabola, S.; Wang, W. cMolGPT: A Conditional Generative Pre-Trained Transformer for Target-Specific De Novo Molecular Generation. *Molecules* **2023**, *28* (11), 4430. https://doi.org/10.3390/molecules28114430.
(43) Zheng, Z.; Zhang, O.; Borgs, C.; Chayes, J. T.; Yaghi, O. M. ChatGPT Chemistry Assistant for Text Mining and the Prediction of MOF Synthesis. *J. Am. Chem. Soc.* **2023**, *145* (32), 18048–18062. https://doi.org/10.1021/jacs.3c05819.
(44) Zheng, Z. AI and Chemistry in Action: Transforming Crystallization for Scalable Water Harvesting Solutions. *ACS Cent. Sci.* **2024**, *10* (12), 2173–2174. https://doi.org/10.1021/acscentsci.4c01838.
(45) Kevlishvili, I.; Michel, R. G. S.; G. Garrison, A.; W. Toney, J.; Adamji, H.; Jia, H.; Román-Leshkov, Y.; J. Kulik, H. Leveraging Natural Language Processing to Curate the tmCAT, tmPHOTO, tmBIO, and tmSCO Datasets of Functional Transition Metal Complexes. *Faraday Discuss.* **2025**, *256* (0), 275–303. https://doi.org/10.1039/D4FD00087K.
(46) Kevlishvili, I.; Duan, C.; Kulik, H. J. Classification of Hemilabile Ligands Using Machine Learning. *J. Phys. Chem. Lett.* **2023**, *14* (49), 11100–11109. https://doi.org/10.1021/acs.jpclett.3c02828.
(47) Toney, J. W.; St. Michel, R. G.; Garrison, A. G.; Kevlishvili, I.; Kulik, H. J. Identifying Dynamic Metal–Ligand Coordination Modes with Ensemble Learning. *J. Am. Chem. Soc.* **2025**, *147* (52), 48218–48234. https://doi.org/10.1021/jacs.5c17169.
(48) Moldagulov, G.; Lee, K.; Nurgaliyev, S.; Salem, A.; Kuznietsov, A.; Grzybowski, B. A. Hybrid Computational Strategy for Predicting Complex Ligand–Metal Architectures. *Angew. Chem. Int. Ed. n/a* (n/a), e24655. https://doi.org/10.1002/anie.202524655.
(49) Kevlishvili, I. T-REX-Ent: Enantiomer Pairs of Chiral TMCs, 2026. https://doi.org/10.5281/zenodo.19103243.
(50) Wu, Z.; Pan, S.; Chen, F.; Long, G.; Zhang, C.; Yu, P. S. A Comprehensive Survey on Graph Neural Networks. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *32* (1), 4–24. https://doi.org/10.1109/TNNLS.2020.2978386.
(51) Brossard, R.; Frigo, O.; Dehaene, D. Graph Convolutions That Can Finally Model Local Structure. arXiv June 3, 2021. https://doi.org/10.48550/arXiv.2011.15069.
(52) Du, W.; Zhang, S.; Cai, Z.; Li, X.; Liu, Z.; Fang, J.; Wang, J.; Wang, X.; Wang, Y. Molecular Merged Hypergraph Neural Network for Explainable Solvation Gibbs Free Energy Prediction. *Research* **2025**, *8*, 0740. https://doi.org/10.34133/research.0740.
(53) Feng, Y.; You, H.; Zhang, Z.; Ji, R.; Gao, Y. Hypergraph Neural Networks. *Proc. AAAI Conf. Artif. Intell.* **2019**, *33* (01), 3558–3565. https://doi.org/10.1609/aaai.v33i01.33013558.

(54) Lin, Y.; Yuan, Q.; Qiu, Q.; Lian, S. 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. https://doi.org/10.1109/ITME63426.2024.00135.
(55) Chen, J.; Schwaller, P. Molecular Hypergraph Neural Networks. *J. Chem. Phys.* **2024**, *160* (14), 144307. https://doi.org/10.1063/5.0193557.
(56) Cho, K.; van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; Bengio, Y. 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)*; Moschitti, A., Pang, B., Daelemans, W., Eds.; Association for Computational Linguistics: Doha, Qatar, 2014; pp 1724–1734. https://doi.org/10.3115/v1/D14-1179.
(57) Zaheer, M.; Kottur, S.; Ravanbakhsh, S.; Poczos, B.; Salakhutdinov, R. R.; Smola, A. J. Deep Sets. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc., 2017; Vol. 30.
(58) Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; Dahl, G. E. Neural Message Passing for Quantum Chemistry. In *Proceedings of the 34th International Conference on Machine Learning*; PMLR, 2017; pp 1263–1272.
(59) van Gerwen, P.; Briling, K. R.; Bunne, C.; Somnath, V. R.; Laplaza, R.; Krause, A.; Corminboeuf, C. 3DReact: Geometric Deep Learning for Chemical Reactions. *J. Chem. Inf. Model.* **2024**, *64* (15), 5771–5785. https://doi.org/10.1021/acs.jcim.4c00104.
(60) Atz, K.; Grisoni, F.; Schneider, G. Geometric Deep Learning on Molecular Representations. *Nat. Mach. Intell.* **2021**, *3* (12), 1023–1032. https://doi.org/10.1038/s42256-021-00418-8.
(61) Batzner, S.; Musaelian, A.; Sun, L.; Geiger, M.; Mailoa, J. P.; Kornbluth, M.; Molinari, N.; Smidt, T. E.; Kozinsky, B. E(3)-Equivariant Graph Neural Networks for Data-Efficient and Accurate Interatomic Potentials. *Nat. Commun.* **2022**, *13* (1), 2453. https://doi.org/10.1038/s41467-022-29939-5.
(62) Ioannidis, E. I.; Gani, T. Z. H.; Kulik, H. J. molSimplify: A Toolkit for Automating Discovery in Inorganic Chemistry. *J. Comput. Chem.* **2016**, *37* (22), 2106–2117. https://doi.org/10.1002/jcc.24437.
(63) Terrones, G.; Michel, R. S.; Toney, J.; Ball, A.; Wang, Y.; Garrison, A.; Nandy, A.; Meyer, R.; Edholm, F.; Oh, C.; Pujet, S.; Chu, D.; Muhammetgulyyev, D.; Kulik, H. molSimplify 2.0: Improved Structure Generation for Automating Discovery in Inorganic Molecular and Reticular Chemistry. ChemRxiv November 19, 2025. https://doi.org/10.26434/chemrxiv-2025-h8gff.
(64) Kevlishvili, I.; Dorabawila, D. Data for Taming T-REX: A Canonical Language for Geometry-Aware Generative Design of Transition-Metal Complexes, 2026. https://doi.org/10.5281/zenodo.19103065.

For Table of Contents Only