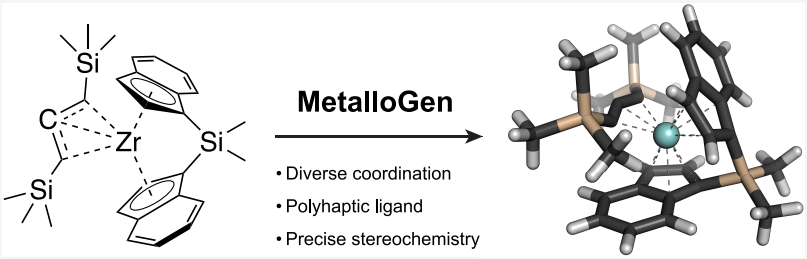
JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Article

MetalloGen: Automated 3D Conformer Generation for Diverse Coordination Complexes

Kyunghoon Lee, Shinyoung Park, Minseong Park, and Woo Youn Kim*

Cite This: https://doi.org/10.1021/acs.jcim.5c02074 Read Online



ABSTRACT: Conformer generation is crucial for computational chemistry tasks such as structure-based modeling and property prediction. Although reliable methods exist for organic molecules, coordination complexes remain challenging due to their diverse coordination geometries, ligand types, and stereochemistry. Current tools often lack the flexibility and reliability required for these systems. Here, we introduce MetalloGen, a novel algorithm designed for the automated generation of 3D conformers of mononuclear coordination complexes. MetalloGen accepts either SMILES strings or molecular graph representations as input and enables the generation of reliable conformers, including those with multiple polyhapto ligands, which are typically inaccessible to conventional conformer generators. To rigorously assess MetalloGen's performance, we benchmarked it on three distinct data sets: a curated collection of experimentally determined structures from the Cambridge Structural Database, the MOR41 benchmark set encompassing a wide range of organometallic reactions and complex ligand environments, and three catalytic reactions. Across all test sets, MetalloGen consistently reproduced appropriate geometries with high fidelity and demonstrated robust stereochemical control, even for challenging cases involving multiple polyhapto ligands. The versatility and reliability of MetalloGen make it a valuable tool for more accurate and efficient computational investigations in inorganic and organometallic chemistry.

1. INTRODUCTION

The generation of 3D molecular conformers is a key step in many computational chemistry workflows.[1,2] Given a molecular graph, conformers are typically constructed using methods such as distance geometry (DG),[3] rule-based approaches like OMEGA,[4] or more recently, machine learning-based methods.[5−12] These initial structures are then refined using force fields (FF), including the Universal Force Field,[13] Merck Molecular Force Field,[14,15] and GFN-FF.[16] For greater accuracy, semiempirical methods such as PM6[17] and GFN-xTB[18−20] or more sophisticated density functional theory (DFT) approaches can be employed. The resulting optimized geometries serve as the foundation for computing molecular and electronic properties, including dipole moments, atomic partial charges, orbital energies, and thermochemical functions. These properties can be used for various applications such as quantitative structure−activity relationship modeling,[21−23] virtual screening,[24,25] machine learning database construction,[26−32] reaction mechanism study,[22,33−39] etc. Therefore, a reliable 3D conformer generation method is a must in computational chemistry workflows.

For typical organic molecules, the generation of conformers is well established. In particular, DG methods such as ETKDG,[2,40] combined with force field optimizations, enable the production of accurate conformers. This approach is widely used in fields such as drug discovery[24,41] and materials discovery.[42−45] However, unlike organic molecules, coordination complexes pose significant challenges for generating 3D conformers due to their structural diversity and complexity.[46−57] These complexes can have a wide range of metal centers spanning the s-, p-, d-, and f-blocks. Moreover, structural diversity is further increased by the various metal−ligand binding modes, including chelation and hapticity. This complexity is compounded by the variability of stereoisomers that can arise within a given coordination environment. As a result, conformer generation strategies developed for organic

Received: August 28, 2025
Revised: October 17, 2025
Accepted: October 20, 2025

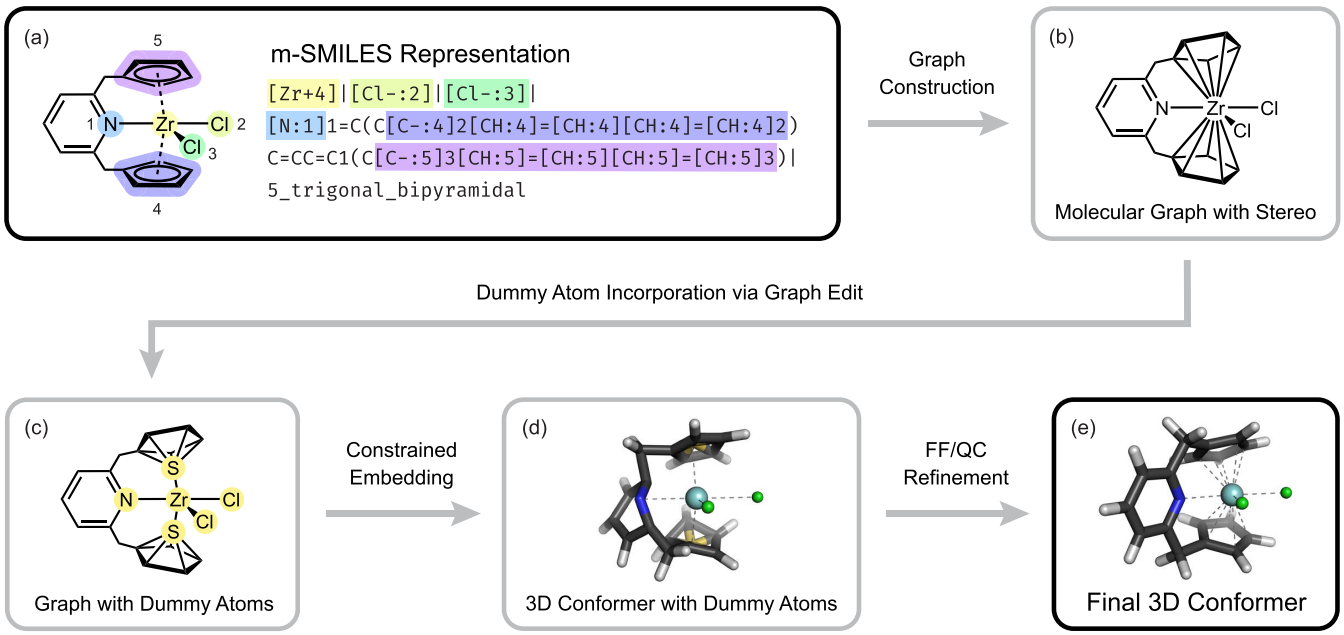


Figure 1. Overview of the MetalloGen algorithm. (a) MetalloGen begins with a modified SMILES (m-SMILES) representation that encodes the metal center (highlighted in yellow), ligands (highlighted in various colors), and overall coordination geometry (trigonal bipyramidal). Each donor atom is labeled with an atom mapping number, enclosed in brackets. This number, placed after a colon (e.g., `[Cl-:2]`), indicates the coordination site within the coordination geometry. (b) From this input, a molecular graph is constructed that includes stereochemistry. Alternatively, this graph can be provided directly by the user. (c) Dummy atoms are added at each polyhapto coordination site to facilitate the handling of polyhapto ligands. (d) A rough 3D structure with the correct stereochemistry is generated using a constrained embedding algorithm. (e) In the final step, dummy atoms are removed, and the structure is refined using force field (FF) or quantum chemistry (QC) methods.

molecules are not directly applicable to coordination complexes.

To address this challenge, several 3D conformer generation tools for coordination complexes have been developed over the past decade.[58−65] One of the earliest tools is MolSimplify, developed by the Kulik group.[58] This method uses rigid body manipulations and force field optimizations to position and orient mono- and bidentate ligands around a metal center. Later, the Reiher group introduced Molassembler,[59] which offers a generalized framework for constructing molecular graphs to enable the detailed classification of coordination geometries and stereochemistry, as well as the generation of conformers. More recent works include Architector,[60] which, for the first time, builds on earlier works to generate conformers of f-block organometallic complexes, and the MACE program,[61] developed to generate all possible stereochemical configurations of octahedral and square planar complexes. In the latest developments, machine learning−based methods, particularly diffusion-based generative models, have been explored for structure generation in coordination complexes as alternatives to algorithm-based approaches.[66−70] The aforementioned methods have been widely used to explore the broad chemical space of transition metal complex catalysts,[71−76] metal−organic frameworks,[73,77,78] molecular magnetic materials,[79−82] and other related systems.[57,83−87] Several recent reviews provide comprehensive overviews of advances in conformer generation for coordination complexes and their applications.[51,76,88,89]

Despite such significant advancements, existing tools still face limitations in the automated generation of 3D conformers for certain classes of coordination complexes, many of which play key roles in coordination chemistry. For example, MolSimplify is not automated for systems containing high-denticity or polyhapto ligands, as it treats ligands as rigid entities. In such cases, ligands must be manually prepared as predefined geometries—referred to as "custom cores"—to enable conformer generation. While Architector can generate 3D conformers of complexes with high-denticity ligands without manual intervention through the DG method, it does not support side-on bound ligands, such as η^2-ethylene in Zeise's salt, and polydentate haptic ligands with donor atoms that do not participate in haptic bonding. Such ligands are especially common in organometallic catalysis, including olefin polymerization, hydrogenation, cross-coupling, and the activation of molecules from dihydrogen to alkanes.[74,90−] Similarly, MACE cannot handle ligands with η interactions and is limited to square planar and octahedral geometries, despite offering unique features such as the systematic enumeration of all feasible stereoisomers. Diffusion-based generative models also do not address these limitations, as these studies have focused on the generative design of novel coordination complexes rather than targeted generation of conformers from a given molecular graph.

In this work, we present a new conformer generation tool called MetalloGen for diverse coordination complexes, to address the aforementioned limitations. Inspired by Architector and MACE, MetalloGen employs the DG method with slight modifications to support side-on bound and polyhapto ligands while enabling precise control over stereochemical configurations. Additionally, it supports conformer generation directly from a SMILES-like representation that encodes the molecular graph and coordination environment, enabling seamless integration with many computational workflows. As a result, MetalloGen can generate conformers across a broad range of coordination geometries and ligand types—including both polydentate and polyhapto ligands—that are commonly

https://doi.org/10.1021/acs.jcim.5c02074
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

encountered in organometallic chemistry and can be readily used for related applications.

To rigorously evaluate MetalloGen's performance, we benchmarked the algorithm against a curated subset of experimentally characterized structures from the Cambridge Structural Database (CSD). These benchmarks demonstrate that MetalloGen reliably generates 3D conformers across diverse coordination environments, highlighting its suitability for high-throughput computational screening of metal-containing compounds, regardless of the application domain or specific use case.

We further assessed MetalloGen using the MOR41 benchmark, which comprises 41 diverse organometallic reactions originally curated by expert computational chemists to evaluate DFT methods.[97] This set includes many side-on bound ligands, such as ethylene and cyclohexene, along with a variety of coordination geometries. Accurate conformer generation for these complexes requires precise control of stereochemistry. Despite this challenge, MetalloGen successfully reproduced all 64 mononuclear organometallic complexes involved in the MOR41 set.

Finally, we applied MetalloGen to three catalytic systems to evaluate its ability to automatically compute reaction energies for elementary steps in each catalytic cycle. In all cases, we found that the resulting energy profiles closely matched the corresponding reference data.

In what follows, we first describe the overall workflow of MetalloGen. Next, we present a detailed evaluation of its performance across the introduced three test sets. Finally, we conclude with a discussion of our findings and future directions for further development.

2. METHODS

f1 Figure 1 illustrates a simplified workflow of our algorithm using a trigonal bipyramidal pyridine-bridged zirconocene dichloride compound as an example. We devised a modified SMILES representation for mononuclear coordination complexes, called m-SMILES, which serves as input for MetalloGen to generate the corresponding 3D conformers (Figure 1a). This representation differs from the SMILES-based format recently developed by Rasmussen et al.,[98] which was designed to be directly parsable by RDKit, enabling seamless integration with cheminformatics tools. Specifically, m-SMILES encodes the metal center (e.g., `[Zr+4]`, highlighted in yellow), the SMILES strings of individual ligands, and the overall coordination geometry (e.g., `5_trigonal_bipyramidal`). Ligands are separated by vertical bars, and their donor atoms directly coordinated to the metal are indicated by square brackets. Coordination sites are assigned using atom mapping numbers; for example, `[Cl-:2]` indicates that a chloro ligand is placed at coordination site 2 (highlighted in light green), and `[C-:4]2[CH:4]=[CH:4][CH:4]=[CH:4]2` specifies that the five carbon atoms of a cyclopentadienyl (Cp) ring are bonded to the metal center at coordination site 4 (highlighted in purple-blue). While this format is not directly RDKit-compatible, it offers a more expressive and flexible syntax that facilitates intuitive specification of coordination geometry and metal-centered stereochemistry. In particular, it enables straightforward encoding of structurally complex ligands, such as polydentate and polyhapto systems like the pyridine-bridged bis-(cyclopentadienyl) ligand illustrated in Figure 1. Starting from this m-SMILES representation, MetalloGen proceeds through four key steps to generate a reliable 3D conformer: molecular graph construction, dummy atom addition, 3D embedding, and structural refinement.

In the first step, a molecular graph is constructed from the given m-SMILES, encoding both atomic connectivity and stereochemical details (Figure 1b). This graph is built by connecting all donor atoms in each ligand and the metal atom. In Figure 1b, edges are added between each carbon atom of the Cp rings and the zirconium center, resulting in a formal valence of 13 for the metal. If this connectivity information is already provided, MetalloGen can bypass this step and directly proceed to 3D conformer generation using the given connectivity.

While this graph can be chemically intuitive, we found that direct 3D embedding of such a high-valence molecular graph often fails when handling it with standard cheminformatics tools (e.g., RDKit). One possible reason for this is the high coordination number of the metal (e.g., 13 for Zr in Figure 1b), which exceeds the typical valency encountered in organic molecules, usually no more than six. To mitigate this issue, MetalloGen introduces dummy atoms at each polyhapto coordination site (Figure 1c). Each Cp ring is now connected to a dummy sulfur atom, which in turn coordinates with the metal center. This effectively reduces the bond count of the zirconium atom from 13 to 5, significantly facilitating the embedding process. This dummy-atom insertion is a key step in our algorithm, allowing reliable 3D embedding of complexes with haptic ligands—regardless of their denticity—while preserving the intended coordination geometry (e.g., trigonal bipyramidal). From this modified graph, MetalloGen proceeds to the 3D embedding step. It employs RDKit's built-in constrained embedding algorithm, which is based on the DG method, to construct an initial 3D structure (Figure 1d). To enforce correct stereochemistry, MetalloGen applies positional constraints derived from predefined coordination templates. These templates consist of sets of normalized direction vectors corresponding to each coordination site, guiding the spatial arrangement of donor atoms directly bonded to the metal center (highlighted in yellow in Figure 1c). In total, 30 templates are implemented, adapted from the Architector toolkit, covering a wide range of coordination environments.

In the final step, the generated structure undergoes structural refinement to restore target metal−ligand distances and optimize ligand geometries (Figure 1e). This step refines distance inaccuracies caused by dummy atoms and other distortions that can incur during embedding. The refinement is carried out using constrained scan optimization, in which geometric constraints are applied to the ligand atoms coordinated to the metal center. The procedure begins with an FF method, and if the FF-based optimization fails to produce a chemically reasonable structure, a quantum chemical (QC) method is applied. In MetalloGen, GFN2-xTB is used as the default QC method, offering a good balance between computational efficiency and chemical accuracy. As a result, MetalloGen produces chemically valid 3D conformers corresponding to the input m-SMILES representation, even for a complex that includes a multidentate haptic ligand. However, the resulting conformers are only partially optimized due to the imposed geometric constraints. Therefore, we note that they should be further relaxed at the desired level of theory prior to downstream applications. For the algorithmic details, including available coordination geometries and

https://doi.org/10.1021/acs.jcim.5c02074
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

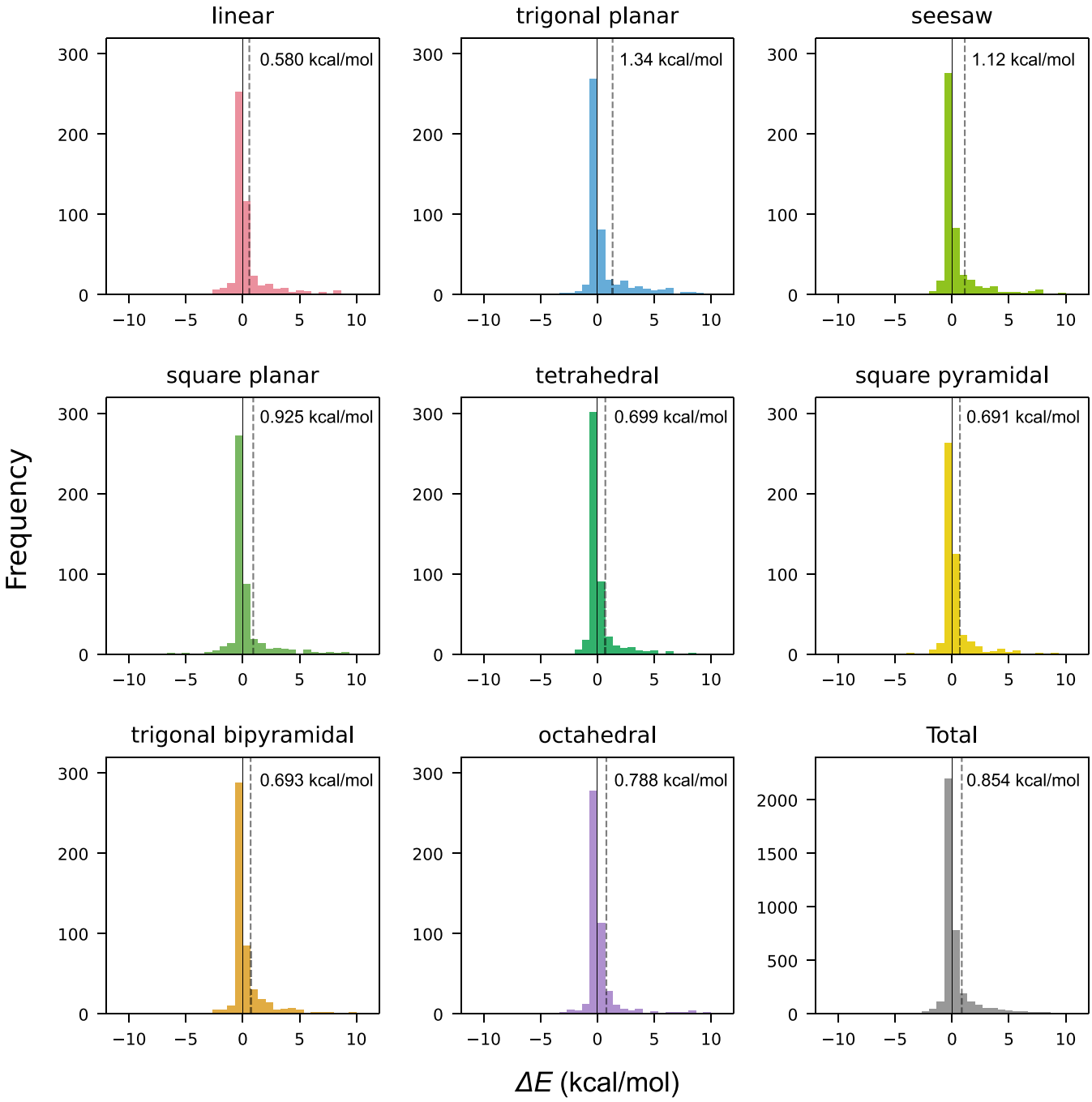


Figure 2. Histogram of energy differences between MetalloGen-generated and their reference structures. For clarity, only samples with absolute energy differences below 10 kcal/mol are shown. The energy difference (ΔE) is defined as the difference between the electronic energies of the MetalloGen and CSD structures ($E_{\text{MetalloGen}} - E_{\text{CSD}}$), both optimized using GFN2-xTB. The solid gray line marks $\Delta E = 0$, and the dashed gray lines indicate the mean ΔE for each coordination geometry, shown in the top right corner.

examples, we refer to Section S1 of the Supporting Information.

3. RESULTS AND DISCUSSION

3.1. The CSD Benchmark Test. We first assessed the reliability of MetalloGen in generating conformers across diverse coordination complexes using a subset of the Cambridge Structural Database (CSD), which contains over 500,000 experimentally validated organometallic complexes.[99] Specifically, we queried version 5.45 of the CSD (June 2024 update) using the CSD Python API to construct a comprehensive benchmark set. We selected mononuclear complexes spanning the eight most frequently observed coordination geometries in the CSD: linear, trigonal planar, square planar, seesaw, tetrahedral, square pyramidal, trigonal bipyramidal, and octahedral. In many computational workflows, generated structures are optimized using QC methods for consistent level of accuarcy. As such, we also applied QC optimizations on the reference CSD structures. This allows us to evaluate the fidelity of MetalloGen-generated geometries against a common computational baseline, rather than comparing them directly to the experimental coordinates.

https://doi.org/10.1021/acs.jcim.5c02074
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

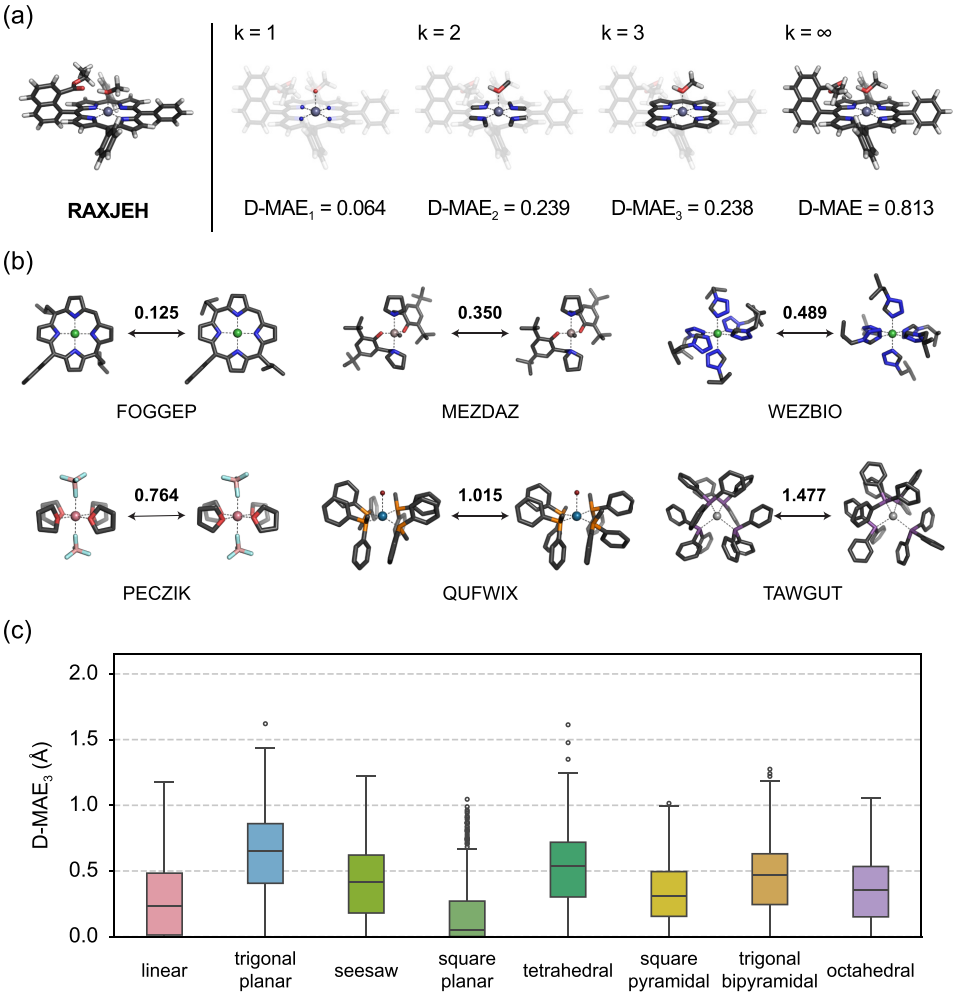


Figure 3. (a) Illustration of D-MAE$_k$ calculations for different values of k. As k increases, additional atoms that are more distant from the metal center are included. (b) Representative examples of complexes with low, medium, and high D-MAE$_3$ values. For each pair, the structure on the left is the CSD reference, and the structure on the right is the struture generated by MetalloGen. (c) Box plot of D-MAE$_3$ values for each coordination geometry.

Considering the substantial computational cost of processing a large number of compounds, we employed the GFN2-xTB method, a cheaper alternative to DFT, for QC optimization. For each geometry type, we sampled 500 different complexes based on several screening criteria, stated as below:

1. **Mononuclear complexes only:** Structures containing a single metal center were selected to simplify the analysis and refinement process.
2. **Moderate molecular size:** Complexes with a total atom count less than or equal to 150 were retained, for a manageable computational cost of QC calculations.
3. **Valence cutoff:** Complexes containing any nonmetal atom with a valence state greater than 4 were excluded, except for the metal center itself.
4. **Successful termination:** No error termination occurred during geometry optimization.
5. **No imaginary frequency:** The Hessian matrix was successfully computed, and no imaginary frequency was present.
6. **Structure preservation:** The assigned coordination geometry (based on shape measures) remained consistent before and after optimization.
7. **Adjacency preservation:** The metal center's adjacency list was preserved, meaning that the set of atoms directly bonded to the metal did not change.

After applying the above criteria, we yielded a benchmark set comprising 4,000 GFN2-xTB−optimized structures. For conformer generation with MetalloGen, up to ten candidate structures were sampled per complex using various hyperparameter settings, as the final relaxed geometries often converge to different local minima depending on the initial configuration. All MetalloGen-generated structures were subsequently reoptimized using the same GFN2-xTB method to ensure consistency. In this evaluation, molecular graph representations were extracted directly from the original CSD SDF files. Additional details on the preparation of both the CSD benchmark set and the MetalloGen-generated structures are provided in Sections S2 and S3 of the Supporting Information.

Among the 4000 test structures, MetalloGen failed to reproduce reference structures in only three cases, achieving a 99.9% success rate. The three cases are shown in Figure S2. In each of the three cases, MetalloGen initially produced chemically plausible conformers, but the final xTB optimization resulted in highly distorted and fragmented structures.

https://doi.org/10.1021/acs.jcim.5c02074
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

When the initial conformers were optimized with DFT at the PBE0-D3(BJ)/def2-SVP level,[100−102] they converged to stable structures that closely matched the DFT-optimized reference structures (see Section S4), confirming the chemical plausibility of the MetalloGen's generated structures. The overall results highlight MetalloGen's exceptional robustness and accuracy across a wide spectrum of organometallic compounds.

Figure 2 shows the energy differences between the structures generated by MetalloGen and their corresponding reference structures. To minimize the effect of conformational variability, we selected the structure with energy closest to the reference. In addition, only samples with absolute energy differences below 10 kcal/mol, which account for 98% of the test set, are shown for clarity. Across all eight geometry types, the overall mean energy difference was 0.854 kcal/mol, with the mean energy differences for each type all below 1.4 kcal/mol. The largest energy difference was observed for the trigonal planar geometry, with a mean energy difference of 1.34 kcal/mol. The mean absolute energy difference across the entire data set was 1.15 kcal/mol, with no geometry type exceeding 2.0 kcal/mol. Notably, about 80% of the structures had energy differences of less than 1 kcal/mol, which is a commonly accepted threshold for chemical accuracy.

Among outliers with energy differences above 10 kcal/mol, most cases were due to variations in hydrogen bonding or differences in the conformations of flexible ring systems in ligands, such as twist-boat versus chair forms. Representative examples of such outliers are shown in Figure S3. These outliers likely arise because such stabilizing effects are not explicitly considered during the embedding step, representing an area for future improvement. Nonetheless, these results demonstrate that MetalloGen can reliably generate 3D conformers across a wide range of coordination geometries.

In addition, we evaluated how the generated conformers structurally differ from the reference by measuring the distance mean absolute error (D-MAE), defined as

$$\text{D-MAE}(D, D') = \frac{1}{N(N-1)} \sum_{i,j<N} |D_{ij} - D'_{ij}| \tag{1}$$

where N is the number of atoms in each structure, D and D' are the interatomic distance matrices of the two given structures. The smaller the D-MAE value, the smaller their structural difference. This metric has been adopted in several studies to assess the performance of structure generation.[103−105] To focus on atoms closer to the metal center, we introduced a localized metric, D-MAE$_k$, defined as follows:

$$\text{D-MAE}_k(D, D') = \frac{1}{|\mathcal{N}_k|(|\mathcal{N}_k| - 1)} \sum_{i,j\in\mathcal{N}_k} |D_{ij} - D'_{ij}| \tag{2}$$

where $\mathcal{N}_k$ denotes the set of atoms within k-nearest neighbors of the metal center. An illustration of D-MAE$_k$ is shown in Figure 3a. As k increases, the D-MAE$_k$ calculation includes more atoms that are farther away from the metal center. For example, when $k = 1$, only atoms directly bonded to the metal are included, resulting in very low D-MAE$_1$ values that underestimate structural deviations. In constrast, when all atoms are included (i.e., $k = \infty$), even minor conformational changes in distant regions of the ligands affect the value. We observed that D-MAE values calculated using all atoms ($k = \infty$) were often substantially higher than those computed with any fixed k values (Figure S4). Upon closer inspection, many of these outliers could be attributed to crystal packing effects in the CSD reference structures, which are derived from solid-state X-ray crystallography. These structures tend to adopt highly symmetric conformations optimized for crystal packing. In contrast, MetalloGen-generated conformers are less symmetric, with flexible ligand groups occasionally forming intramolecular interactions with the metal center. Representative examples are provided in Figure S5. Notably, in these examples, although the D-MAE$_\infty$ values ranged from 3 to 6 Å, the corresponding D-MAE$_3$ values were significantly lower (less than 0.6 Å), and the electronic energy differences remained small. This indicates that the core coordination environment around the metal center was well preserved despite deviations in the outer ligand conformations. Based on these findings, we selected $k = 3$ for evaluation in this study, as it captures the chemically relevant region as effectively as possible near the metal center while minimizing the influence of peripheral conformational differences.

Figure 3b shows six representative examples of D-MAE$_3$ ranging from 0.1 to 1.5 Å. In the cases with the smallest deviations (CSD refcode FOGGEP, 0.125 Å), the MetalloGen- and CSD-derived structures are nearly superimposable. For structures with D-MAE$_3$ values around 0.4 Å (MEZDAZ, 0.350 Å; WEZBIO, 0.489 Å), only minor conformational differences were observed, such as slight distortions in rings or small rotations of distal methyl groups. For larger D-MAE$_3$ values, the structures remained chemically equivalent to their references but exhibited different degrees of deviation depending on the conformational variability of the ligands. In PECZIK (0.764 Å), the deviation arose from the rotation of an axial tetrafluoroborate ligand. In QUFWIX (1.015 Å), the large deviation was caused by the rotation of two bulkier dicyclohexyl(methyl)phosphine ligands in the equatorial plane. The largest deviation was observed in TAWGUT (1.477 Å), which contains four tris(cyclohexyl)stibine ligands. The large size and flexibility of these stibine ligands, combined with the relatively large atomic radius of the coordinating antimony (Sb) atom, caused the highest D-MAE$_3$.

Figure 3c presents the distribution of D-MAE$_3$ values for each coordination geometry as a box plot, with the same set of CSD complexes used in the energy comparison of Figure 2. Most coordination geometries exhibited average D-MAE$_3$ values below 0.5 Å, indicating that MetalloGen- and CSD-derived structures differ only slightly due to minor conformational variations, which is consistent with the low energy differences observed earlier. The overall average D-MAE$_3$ across all geometries was 0.399 Å, with the smallest deviation of 0.173 Å observed for the square planar geometry. While most D-MAE values fall within a small range, a few notable outliers were found in the trigonal planar, square planar, tetrahedral, and trigonal bipyramidal geometries. In particular, the square planar geometry had more outliers, ranging from 0.7 to 1.1 Å. For the remaining geometries, a few outliers with D-MAE$_3$ values ranging from 1.0 to 1.6 Å were found, comparable to the deviation observed for TAWGUT in Figure 3b. Manual inspection revealed that these large deviations were primarily caused by significant conformational rearrangements of ligands directly coordinated to the metal center. Both energetic and structural analyses suggest that MetalloGen can generate 3D structures that closely resemble their references across all geometry types considered in this study.

Finally, we examined the diversity of ligands in the test set by analyzing three key properties: the total number of atoms,

https://doi.org/10.1021/acs.jcim.5c02074
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

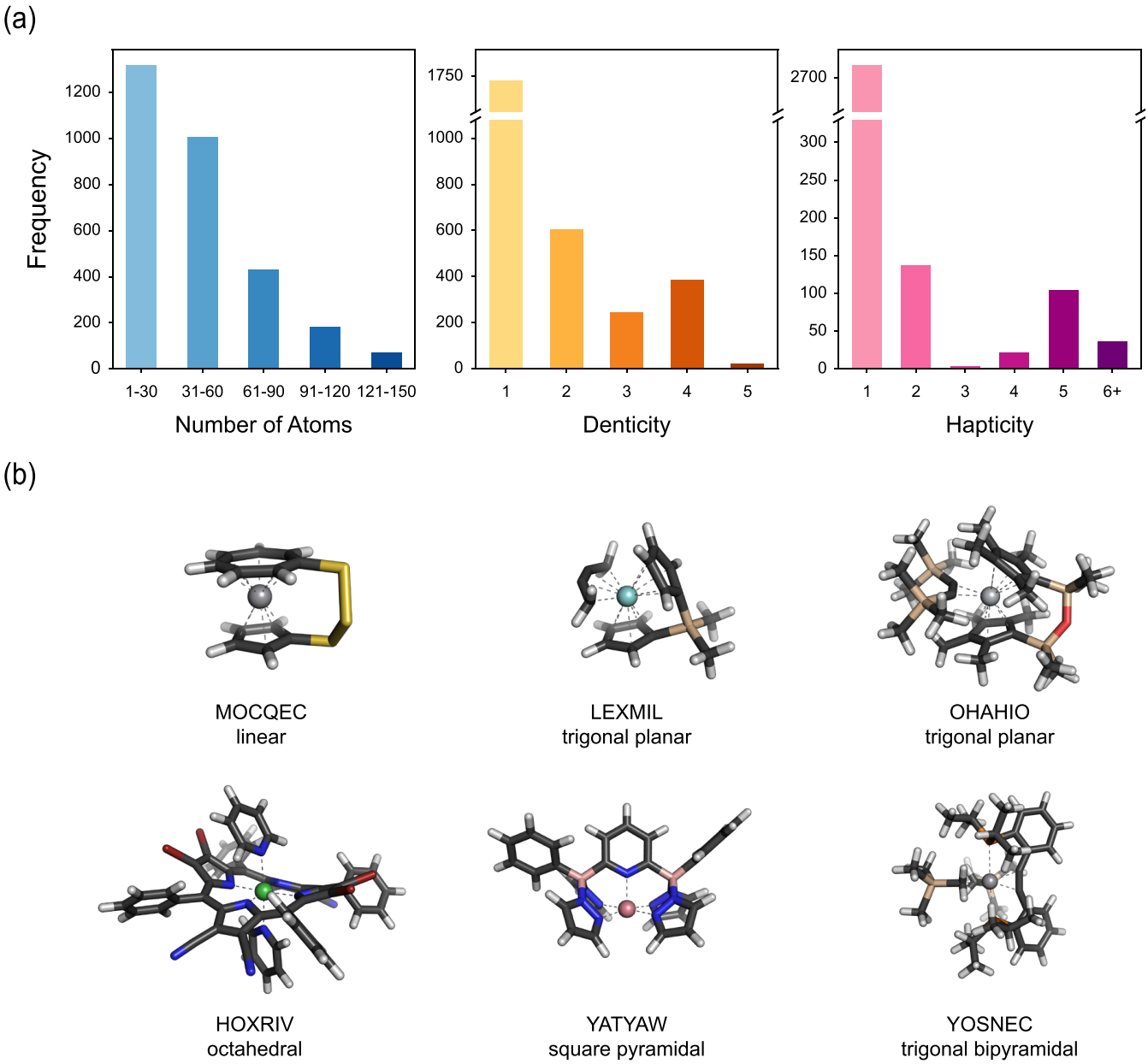


Figure 4. Ligand diversity of the CSD structures used in the benchmark. (a) Distribution of ligand properties, including the total number of atoms, denticity, and hapticity. (b) Representative examples of structures containing multidentate and polyhapto ligands. The top row shows examples with polyhapto ligands such as Cp rings and π-donating groups, while the bottom row shows structures with multidentate ligands.

denticity, and hapticity. Figure 4a shows the distribution of these properties. First, the results show a wide range of ligand sizes in the test set. Ligands with fewer than 30 atoms were the most common, but several hundred had more than 30 atoms, and dozens contained between 121 and 150 atoms. Furthermore, the test set contained a diverse set of structurally complex ligands, including multidentate and polyhapto species, which are the very cases that our work aims to address. While most of the ligands were monodentate or nonhapto, the set included over 1000 multidentate ligands and more than 200 polyhapto ligands. The most common hapticities among the polyhapto ligands were η^2 and η^5, corresponding to well-established organometallic motifs. The η^2 hapticity typically involves side-on binding of σ bonds, such as in H_2 and C−H bonds in σ-complexes, as well as side-on binding of π bonds, such as in alkenes (e.g., ethylene) in π-complexes. The η^5 hapticity is characteristic of conjugated five-membered rings such as the cyclopentadienyl ligand. MetalloGen successfully generated valid structures for these diverse cases, underscoring its ability to handle a wide range of coordination complexes with structurally diverse and complex ligands.

Figure 4b shows examples of the most complex ligands, successfully generated by MetalloGen. The three examples shown at the top of Figure 4b−MOCQEC, LEXMIL, and OHAHIO−contain polyhapto ligands. MOCQEC is a bridged sandwich complex where a single ligand features two aromatic rings simultaneously coordinating to the metal center. LEXMIL is a trigonal planar *ansa*-metallocene composed of a bridged bidentate ligand with two Cp rings and a monodentate ligand bound to the metal center through a linear η^4 interaction. OHAHIO is another trigonal planar *ansa*-metallocene that features a bidentate ligand with two Cp rings linked

https://doi.org/10.1021/acs.jcim.5c02074
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

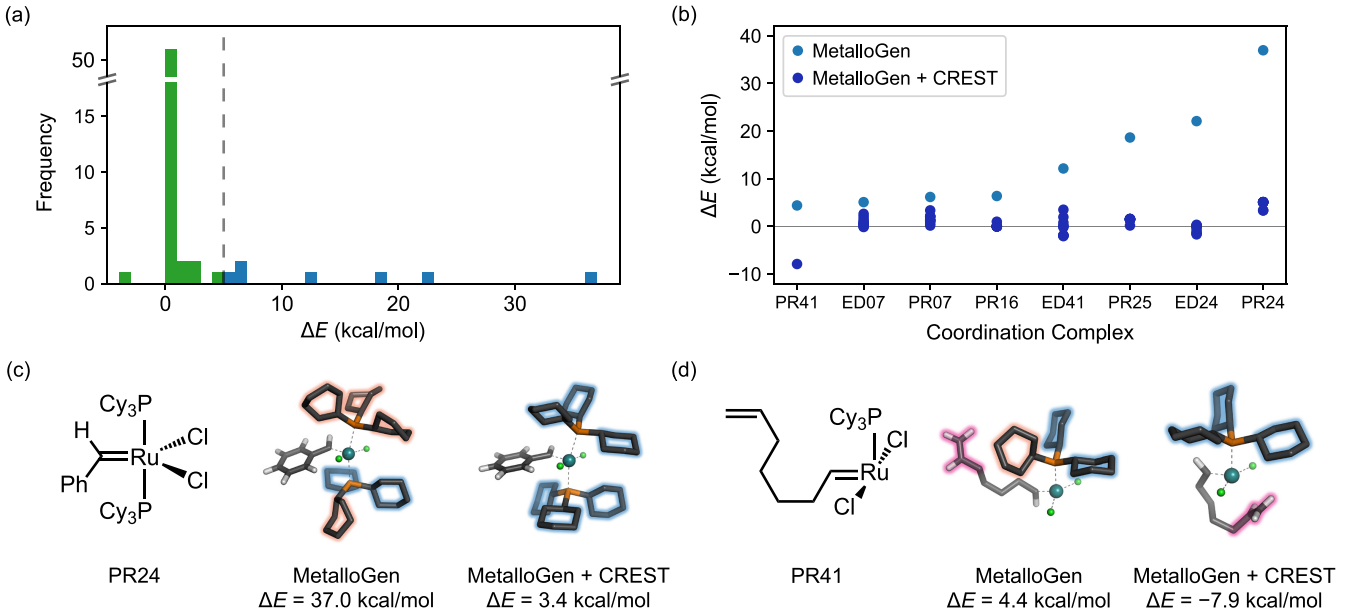


Figure 5. Testing MetalloGen on the MOR41 benchmark set, comprising 64 organometallic complexes derived from 38 reactions. (a) Histogram showing energy differences of MetalloGen-generated structures from their references, which were optimized at the PBE0-D3(BJ)/def2-SVP level. Blue bars on the right side of the dashed line represent structures with energy differences exceeding 5 kcal/mol, which were further refined using CREST. (b) The energy differences of the eight structures above 5 kcal/mol without CREST (sky blue) and those after applying CREST (deep blue). For each coordination complex, the ten lowest-energy conformers identified by CREST were individually reoptimized at the same DFT level. (c−d) PR24 and PR41 conformers generated by MetalloGen, with and without CREST. Cyclohexyl groups in twist-boat conformations are highlighted in red; chair conformations are highlighted in blue. In the PR41 conformer generated by MetalloGen and CREST, the carbon−carbon double bond weakly interacts with the Ru metal center (C−Ru distance of 2.9 Å, highlighted in pink), further stabilizing the conformer.

by a three-atom bridge, along with a monodentate ligand coordinating through an η^2 interaction. The bottom three examples (HOXRIV, YATYAW, and YOSNEC) feature exceptionally high ligand denticity. HOXRIV is an octahedral complex with a tetradentate porphyrin ring in the equatorial position and two axially bound pyridine ligands. YATYAW is a square pyramidal complex with a pentadentate ligand, the highest denticity observed in our test set. YOSNEC is a trigonal bipyramidal complex with a tridentate ligand that includes a single η^2-binding site at one of the equatorial positions. These examples highlight the complexity of generating 3D conformers for the CSD test set and demonstrate MetalloGen's robustness in handling such challenging cases.

3.2. The MOR41 Benchmark Set. The MOR41 benchmark set contains 41 closed-shell organometallic reactions representative of key chemical transformations commonly found in transition metal chemistry and catalysis, including complexation, oxidative addition, and ligand exchange.[97] Compared to the CSD data set, MOR41 offers a more realistic benchmark for practical applications and allows for the evaluation of MetalloGen's potential for automated analysis of organometallic reactions. Moreover, these reactions often involve complexes with η interactions and require accurate generation of structures with precise stereochemistry, where existing methods are likely to struggle. Of the 41 reactions, those involving polynuclear complexes were excluded, yielding a final set of 38 reactions comprising 64 mononuclear organometallic compounds. The original benchmark data set was provided at the DLPNO−CCSD(T)/CBS(def2-TZVPP/def2-QZVPP) level of theory, which is prohibitively expensive. To make the study computationally feasible, we reoptimized all reference geometries and evaluated their energies at the PBE0-D3(BJ) level of theory.[100,101] For these calculations we used the def2-SVP basis set and corresponding effective core potentials,[102] as obtained from the Basis Set Exchange.[106] Then, we prepared the m-SMILES representations for each complex (See Figure S6 for examples) and used them as input to MetalloGen to regenerate their 3D structures. The resulting geometries were also optimized at the PBE0-D3(BJ)/def2-SVP level to ensure consistency. All structures were validated via vibrational frequency analysis. Geometry optimizations and frequency calculations were performed using Gaussian 16.[107] Additional computational details can be found in Section S8.

Figure 5a presents the energy differences between MetalloGen-generated structures and their corresponding reference structures. Of the 64 structures analyzed, 56 had energy differences of less than 5 kcal/mol, and 51 of these were almost identical to the reference. These results manifest the forte of MetalloGen in generating accurate 3D structures of organometallic complexes frequently observed in practical applications. The remaining eight structures with energy differences above 5 kcal/mol exhibited conformationally flexible ligands, such as tricyclohexylphosphine (PCy_3; PR41, ED41, ED24, PR24), triisopropylphosphine ($P(i\text{-}Pr)_3$; ED07, PR07), and 1,3-bis(2,4,6-trimethylphenyl)imidazole (SIMes; PR16, PR25). These bulky ligands with large torsional degrees of freedom led to energy differences as high as 37.0 kcal/mol (Figure 5c, PR24 with two PCy_3 ligands).

To examine whether the observed energy discrepancies can be resolved through additional conformational sampling, we applied the CREST algorithm to the eight structures.[108] Positional constraints were imposed on the metal center and all donor atoms to preserve its stereochemistry. The GFN2-xTB method was used for the CREST sampling, and the resulting ten lowest-energy conformers were subsequently

https://doi.org/10.1021/acs.jcim.5c02074
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

reoptimized with DFT. Additional details regarding the CREST sampling are provided in Section S9. As shown in Figure 5b, conformers with energies comparable to those of their respective references were obtained for all eight complexes. The most dramatic decrease in energy was observed for PR24, where four of the six cyclohexyl groups (Figure 5c, highlighted in orange) switched from an unstable twist-boat conformation to a more favorable chair conformation (Figure 5c, highlighted in blue). In this new conformer, with all six cyclohexyl rings adopting the chair conformation, the energy is 40.4 kcal/mol lower than that of the previously obtained structure and only 3.4 kcal/mol higher than the reference. Interestingly, for PR41, we identified a new conformer lying 12.3 kcal/mol below its earlier counterpart and 7.9 kcal/mol below the reference structure. This substantial stabilization comes from the favorable conformational switch from twist-boat to chair, along with an additional interaction between the Ru center and a nearby carbon–carbon double bond (Figure 5d). These results demonstrate that combining MetalloGen with a conventional conformational sampling tool like CREST can yield low-energy structures suitable for computational studies of organometallic reactions.

Building upon the high reliability of MetalloGen in adjusting the stereochemistry of coordination complexes, we further evaluated whether it can systematically enumerate all feasible stereoisomers of a given complex. As a case study, we selected the PR08 complex from the MOR41 benchmark set, an octahedral Ir(III) complex featuring four different ligand types, including two identical triphenylphosphine and two hydride ligands—providing a suitable test case for stereochemical variation. We enumerated all possible stereoisomeric configurations arising from ligand permutations across the six coordination sites and used MetalloGen to generate the corresponding 3D structures. Each structure was subjected to the CREST algorithm to identify low-energy conformations, and the lowest-energy conformer for each stereoisomer was subsequently refined using DFT at the PBE0-D3(BJ)/def2-SVP level of theory.

Figure 6 shows the eight stereoisomers successfully generated by MetalloGen. The resulting structures displayed a range of relative electronic energies, implying distinct stereoisomeric configurations compared to the original complex. Among them, two enantiomeric pairs (isomers 5–6 and 7–8) were identified, each exhibiting nearly identical electronic energies (differences less than 1 kcal/mol), consistent with mirror symmetry. Notably, two stereoisomers (isomers 4 and 7) were found to be more stable than the original configuration. These results demonstrate that MetalloGen can be effectively used to systematically explore metal-centered stereoisomerism, enabling the identification of more stable or catalytically relevant configurations in coordination complexes.

3.3. Application to Mechanistic Studies of Organometallic Catalysis. The final test set consists of three catalytic reactions characterized by distinct coordination geometries. The first example is a Rh(III)-catalyzed direct C–H amination involving a pentamethylcyclopentadienyl (Cp*) ligand, studied by Park et al.[109] The second is a room-temperature Cu-catalyzed aryl bromide amination, developed by Kim et al.[110] The third is a hydroaryloxylation of an olefin catalyzed by a pincer iridium complex, reported by Haibach et al.[111] These systems were selected to cover a broad range of coordination

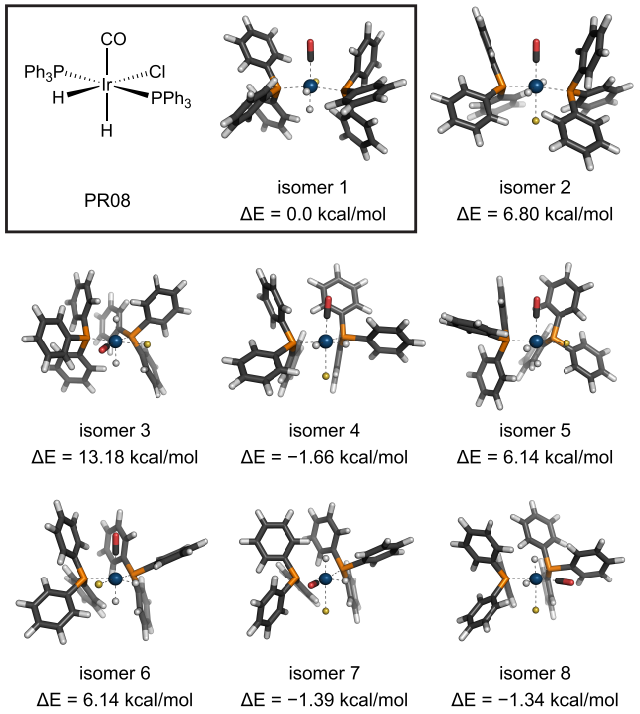


Figure 6. Demonstration of MetalloGen's ability to systematically generate all feasible stereoisomers. The PR08 complex (from the MOR41 benchmark set) is an octahedral Ir(III) complex featuring four distinct ligand types: two triphenylphosphine ligands, two hydrides, one chloride, and one carbonyl. All eight theoretically possible stereoisomers were successfully generated. Each structure was optimized at the PBE0-D3(BJ)/def2-SVP level of theory, and relative electronic energies (in kcal/mol) are reported with respect to the original configuration (isomer 1).

environments: Park et al. includes trigonal planar and tetrahedral geometries; Kim et al. features predominantly square planar geometries; and Haibach et al. exhibits various geometries such as square planar, square pyramidal, and octahedral.

Compared to the previous benchmarks, these catalytic reactions impose additional challenges, including high-energy intermediates (e.g., the Rh(V) nitrenoid species in Park et al.), sterically hindered ligands (e.g., N^1, N^2-diarylbenzene-1,2-diamine ligands of Kim et al.), and stereochemical requirements critical for regioselective outcomes (e.g., Markovnikov-type addition in Haibach et al.). As MetalloGen is designed to provide reasonable initial guesses for local minima structures, we focus on reaction energy calculations (energies of intermediates for each elementary step), leaving activation energies and transition state characterization for future work. To generate the 3D structures of the intermediates, CREST was used to sample low-energy conformers. The lowest-energy conformer for each structure was then reoptimized using DFT. The DFT calculations were performed following the computational protocols outlined in the original studies. More details can be found in Sections S8 and S9 of the Supporting Information.

Figure 7 shows the reaction energy profiles obtained using MetalloGen, alongside the reference energy profile reported in the original studies. MetalloGen successfully reproduced the energy profiles along all three catalytic cycles, with most structures differing by less than 3 kcal/mol and none deviating more than 5 kcal/mol from the reference values. Structural

https://doi.org/10.1021/acs.jcim.5c02074
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

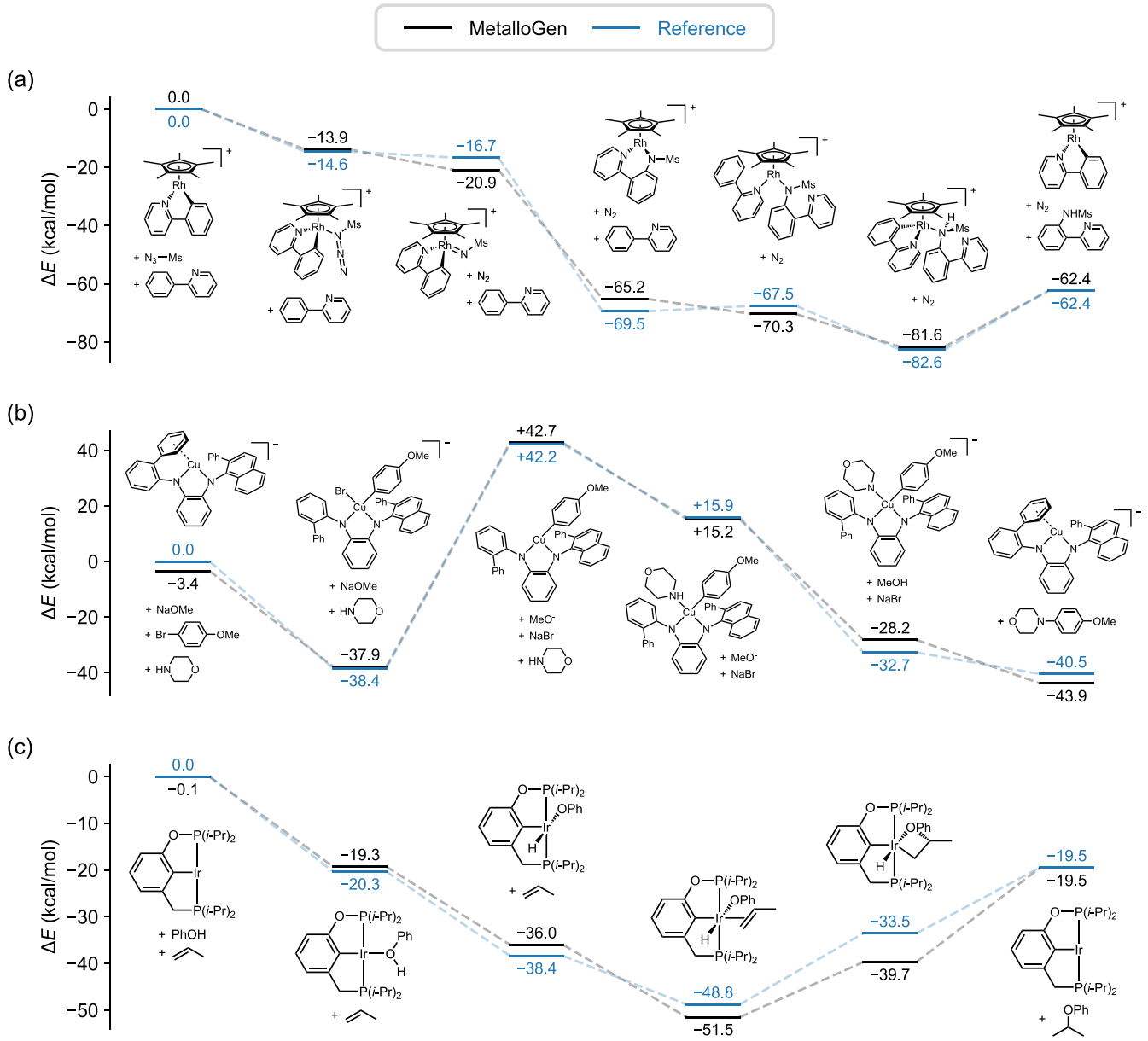


Figure 7. Energy profiles of the three catalytic reactions obtained with MetalloGen combined with CREST. The reference energies are shown in blue, and the energies from MetalloGen-generated structures are shown in black. All energies are given relative to the first reference structure in each cycle. (a) The reaction energy profile of the C−H amination of 2-phenylpyridine with methanesulfonyl azide and a Cp*Rh(III) catalyst by Park et al.[109] (b) The reaction energy profile of the C−N coupling of 4-bromoanisole and morpholine, catalyzed by a diamine-Cu complex, by Kim et al.[110] (c) The reaction energy profile of the propene hydroaryloxylation catalyzed by a pincer-Ir complex, by Haibach et al.[111].

analysis reveals that these energy differences are mainly due to conformational variations. In particular, for the pincer-Ir system, MetalloGen identified an intermediate conformer that is significantly more stable than the one reported in the reference. This intermediate arises from a 1,2-addition of the Ir−O bond to the double bond of the η^2-coordinated propene, forming a four-membered ring with new Ir−C and C−O bonds (Figure 7c, the fifth intermediate). The new conformer is 6.2 kcal/mol more stable than the corresponding reference structure. These results demonstrate MetalloGen's ability to reliably reproduce, and even in some cases, improve the reference structures.

4. CONCLUSIONS AND OUTLOOK

Generating the 3D conformers of coordination complexes is a crucial step in computational workflows for studying metal coordination complexes. While existing methods have made substantial progress, they remain limited in handling complexes with side-on bound and polyhapto ligands, which are commonly encountered in organometallic chemistry. To address these gaps, we developed MetalloGen, a new conformer generation method that supports a wide range of coordination geometries, ligand types, and stereochemical configurations. MetalloGen was evaluated on a curated subset of CSD structures encompassing eight commonly observed coordination geometries. The results show that MetalloGen reliably generates chemically valid conformers across a wide

https://doi.org/10.1021/acs.jcim.5c02074
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

range of ligands, varying in both denticity and hapticity, under diverse coordination environments. Building on this robustness, we applied MetalloGen to the MOR41 benchmark set and three catalytic reactions, successfully reproducing the structures of all mononuclear species involved in these reactions. This enabled a fully automated workflow to calculate reaction energy profiles across a diverse set of organometallic reactions. When coupled with CREST, MetalloGen can yield structures with electronic energies comparable to or lower than those of the reference structures. In addition, MetalloGen supports a SMILES-like input format, termed m-SMILES, which enables users to represent diverse coordination complexes and directly generate their 3D structures. Overall, MetalloGen offers an efficient and automated solution for generating 3D structures of coordination complexes with minimal manual intervention, particularly those relevant to organometallic catalysis.

Despite these advancements, several limitations remain. First, MetalloGen currently supports only mononuclear complexes. As a result, three out of 41 reactions in the MOR41 benchmark set involving binuclear species were excluded from this study. Extending MetalloGen to support polynuclear systems would expand its applicability to a broader range of coordination environments, including those commonly found in multinuclear metalloenzymes, catalysts, and other functional materials.[112−114] Second, MetalloGen shows a higher failure rate for complexes with high coordination numbers (typically seven or more). This limits its applicability to lanthanide and actinide complexes, where alternative tools such as Architector may be more appropriate, although these tools still face limitations when dealing with side-on or polyhapto ligands. Lastly, MetalloGen does not guarantee generation of the lowest-energy conformers. As shown earlier, some generated structures displayed higher energies due to the absence of stabilizing features such as hydrogen bonding or favorable ring conformations. While subsequent conformational refinement using tools like CREST can alleviate these issues, such procedures entail significant computational cost.

The last two limitations primarily arise from the RDKit-based embedding step. This step often fails for complexes with high coordination numbers, thereby interrupting the subsequent steps in MetalloGen. Moreover, the embedding algorithm lacks chemical awareness of subtle stabilizing interactions, which can lead to the generation of higher-energy conformers. Future work could advance in several directions. One is the development of metal-aware distance geometry embedding algorithms to improve the success rate of conformer generation for complexes with high coordination numbers. Another promising direction is the integration of machine learning approaches, particularly diffusion-based generative models, as a means to directly generate low-energy conformers without relying on exhaustive sampling. Such capabilities have already been demonstrated in prior studies on organic molecular systems.[5−12] With sufficient data augmentation using MetalloGen, these strategies could be extended to coordination complexes. Nevertheless, the current version of MetalloGen provides a practical and effective solution that complements existing tools for high-throughput screening and automated mechanistic studies in coordination chemistry, serving as a solid foundation for future computational workflows.

■ ASSOCIATED CONTENT

Data Availability Statement

The raw output log files, including optimized structures and vibrational frequency information, are available in ref 115. The source code for this study is available at https://github.com/kyunghoonlee777/MetalloGen

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.5c02074.

Implementation details and hyperparameters for MetalloGen, CSD preparation details, GFN2-xTB failure modes, outliers in CSD replication, D-MAE$_k$ box plots for various k values, m-SMILES examples for the MOR41 set, DFT and CREST calculation details (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Woo Youn Kim − *Department of Chemistry, KAIST, Daejeon 34141, Republic of Korea; InnoCORE AI-CRED Institute, Korea Advanced Institute of Science and Technology (KAIST), Daejeon 34141, Republic of Korea;* orcid.org/0000-0001-7152-2111; Email: wooyoun@kaist.ac.kr

Authors

Kyunghoon Lee − *Department of Chemistry, KAIST, Daejeon 34141, Republic of Korea; InnoCORE AI-CRED Institute, Korea Advanced Institute of Science and Technology (KAIST), Daejeon 34141, Republic of Korea;* orcid.org/0000-0002-5454-595X

Shinyoung Park − *Department of Chemistry, KAIST, Daejeon 34141, Republic of Korea;* orcid.org/0009-0000-3648-5124

Minseong Park − *Department of Chemistry, KAIST, Daejeon 34141, Republic of Korea*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.5c02074

Author Contributions

K.L. developed the algorithm and designed the study. M.P. and S.P. conducted the experiments, analyzed the results, and made improvements to the algorithm. K.L., S.P., and M.P. wrote the manuscript. W.Y.K. contributed to writing and revising the manuscript.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by the Korea Environmental Industry and Technology Institute (Grant No. RS202300219144) to K.L. and W.Y.K.; Samsung Electronics Co., Ltd. (Grant No. IO231103-07636-01) to K.L. and W.Y.K.; and the InnoCORE program of the Ministry of Science and ICT (Grant No. N10250153) to K.L. and W.Y.K. We also thank Prof. Mu-Hyun Baik of the Center for Catalytic Hydrocarbon Functionalizations, Institute for Basic Science (IBS), and the Department of Chemistry at KAIST for providing computing resources.

REFERENCES

(1) Hawkins, P. C. D. Conformation generation: the state of the art. *J. Chem. Inf. Model.* **2017**, *57*, 1747−1756.

(2) Riniker, S.; Landrum, G. A. Better informed distance geometry: using what we know to improve conformation generation. *J. Chem. Inf. Model.* **2015**, *55*, 2562−2574.

(3) Blaney, J. M.; Dixon, J. S. Distance geometry in molecular modeling. *Rev. Comput. Chem.* **1994**, *5*, 299−335.

(4) Hawkins, P. C. D.; Nicholls, A. Conformer generation with OMEGA: learning from the data set and the analysis of failures. *J. Chem. Inf. Model.* **2012**, *52*, 2919−2936.

(5) Mansimov, E.; Mahmood, O.; Kang, S.; Cho, K. Molecular geometry prediction using a deep generative graph neural network. *Sci. Rep.* **2019**, *9*, No. 20381.

(6) Luo, S.; Shi, C.; Xu, M.; Tang, J. Predicting molecular conformation via dynamic graph score matching. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 19784−19795.

(7) Shi, C.; Luo, S.; Xu, M.; Tang, J.Learning gradient fields for molecular conformation generation *International Conference on Machine Learning* 2021, pp 9558−9568.

(8) Xu, M.; Luo, S.; Bengio, Y.; Peng, J.; Tang, J. Learning Neural Generative Dynamics for Molecular Conformation Generation *International Conference on Learning Representations*, 2021.

(9) Ganea, O.; Pattanaik, L.; Coley, C.; Barzilay, R.; Jensen, K.; Green, W.; Jaakkola, T. Geomol: Torsional geometric generation of molecular 3d conformer ensembles. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 13757−13769.

(10) Xu, M.; Yu, L.; Song, Y.; Shi, C.; Ermon, S.; Tang, J. GeoDiff: A Geometric Diffusion Model for Molecular Conformation Generation *International Conference on Learning Representations*, 2022.

(11) Zhang, H.; Li, S.; Zhang, J.; Wang, Z.; Wang, J.; Jiang, D.; Bian, Z.; Zhang, Y.; Deng, Y.; Song, J.; et al. SDEGen: learning to evolve molecular conformations from thermodynamic noise for conformation generation. *Chem. Sci.* **2023**, *14*, 1557−1568.

(12) Xu, C.; Deng, X.; Lu, Y.; Yu, P. Generation of molecular conformations using generative adversarial neural networks. *Digital Discovery* **2025**, *4*, 161−171.

(13) Rappé, A. K.; Casewit, C. J.; Colwell, K.; Goddard, W. A., III; Skiff, W. M. UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations. *J. Am. Chem. Soc.* **1992**, *114*, 10024−10035.

(14) Halgren, T. A. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *J. Comput. Chem.* **1996**, *17*, 490−519.

(15) Halgren, T. A. Merck molecular force field. II. MMFF94 van der Waals and electrostatic parameters for intermolecular interactions. *J. Comput. Chem.* **1996**, *17*, 520−552.

(16) Spicher, S.; Grimme, S. Robust atomistic modeling of materials, organometallic, and biochemical systems. *Angew. Chem., Int. Ed.* **2020**, *59*, 15665−15673.

(17) Stewart, J. J. P. Optimization of parameters for semiempirical methods V: Modification of NDDO approximations and application to 70 elements. *J. Mol. Model.* **2007**, *13*, 1173−1213.

(18) Grimme, S.; Bannwarth, C.; Shushkov, P. A robust and accurate tight-binding quantum chemical method for structures, vibrational frequencies, and noncovalent interactions of large molecular systems parametrized for all spd-block elements (Z= 1−86). *J. Chem. Theory Comput.* **2017**, *13*, 1989−2009.

(19) Bursch, M.; Hansen, A.; Grimme, S. Fast and reasonable geometry optimization of lanthanoid complexes with an extended tight binding quantum chemical method. *Inorg. Chem.* **2017**, *56*, 12485−12491.

(20) Bannwarth, C.; Ehlert, S.; Grimme, S. GFN2-xTB—An accurate and broadly parametrized self-consistent tight-binding quantum chemical method with multipole electrostatics and density-dependent dispersion contributions. *J. Chem. Theory Comput.* **2019**, *15*, 1652−1671.

(21) Muratov, E. N.; Bajorath, J.; Sheridan, R. P.; Tetko, I. V.; Filimonov, D.; Poroikov, V.; Oprea, T. I.; Baskin, I. I.; Varnek, A.; Roitberg, A.; et al. QSAR without borders. *Chem. Soc. Rev.* **2020**, *49*, 3525−3564.

(22) Crawford, J. M.; Kingston, C.; Toste, F. D.; Sigman, M. S. Data science meets physical organic chemistry. *Acc. Chem. Res.* **2021**, *54*, 3136−3148.

(23) Tropsha, A.; Isayev, O.; Varnek, A.; Schneider, G.; Cherkasov, A. Integrating QSAR modelling and deep learning in drug discovery: the emergence of deep QSAR. *Nat. Rev. Drug Discovery* **2024**, *23*, 141−155.

(24) McNutt, A. T.; Bisiriyu, F.; Song, S.; Vyas, A.; Hutchison, G. R.; Koes, D. R. Conformer generation for structure-based drug design: how many and how good? *J. Chem. Inf. Model.* **2023**, *63*, 6598−6607.

(25) Ree, N.; Koerstz, M.; Mikkelsen, K. V.; Jensen, J. H. Virtual screening of norbornadiene-based molecular solar thermal energy storage systems using a genetic algorithm. *J. Chem. Phys.* **2021**, *155*, No. 184105, DOI: 10.1063/5.0063694.

(26) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; Von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Sci. Data* **2014**, *1*, No. 140022.

(27) Smith, J. S.; Isayev, O.; Roitberg, A. E. ANI-1, A data set of 20 million calculated off-equilibrium conformations for organic molecules. *Sci. Data* **2017**, *4*, No. 170193.

(28) Levine, D. S.; Shuaibi, M.; Spotte-Smith, E. W. C.; Taylor, M. G.; Hasyim, M. R.; Michel, K.; Batatia, I.; Csányi, G.; Dzamba, M.; Eastman, P. et al. *The Open Molecules 2025 (OMol25) Dataset, Evaluations, and Models*, arXiv:2505.08762. arXiv.org e-Print archive. https://arxiv.org/abs/2505.08762, **2025**.

(29) Anstine, D. M.; Zubatyuk, R.; Isayev, O. AIMNet2: a neural network potential to meet your neutral, charged, organic, and elemental-organic needs. *Chem. Sci.* **2025**, *16*, 10228−10244.

(30) Smith, J. S.; Zubatyuk, R.; Nebgen, B.; Lubbers, N.; Barros, K.; Roitberg, A. E.; Isayev, O.; Tretiak, S. The ANI-1ccx and ANI-1x data sets, coupled-cluster and density functional theory properties for molecules. *Sci. Data* **2020**, *7*, No. 134.

(31) Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine learning for molecular and materials science. *Nature* **2018**, *559*, 547−555.

(32) Fabrizio, A.; Grisafi, A.; Meyer, B.; Ceriotti, M.; Corminboeuf, C. Electron density learning of non-covalent systems. *Chem. Sci.* **2019**, *10*, 9424−9432.

(33) Cheong, P. H.-Y.; Legault, C. Y.; Um, J. M.; Celebi-Olcum, N.; Houk, K. Quantum mechanical investigations of organocatalysis: mechanisms, reactivities, and selectivities. *Chem. Rev.* **2011**, *111*, 5042−5137.

(34) Reid, J. P.; Sigman, M. S. Comparing quantitative prediction methods for the discovery of small-molecule chiral catalysts. *Nat. Rev. Chem.* **2018**, *2*, 290−305.

(35) Xie, X.; Clark Spotte-Smith, E. W.; Wen, M.; Patel, H. D.; Blau, S. M.; Persson, K. A. Data-driven prediction of formation mechanisms of lithium ethylene monocarbonate with an automated reaction network. *J. Am. Chem. Soc.* **2021**, *143*, 13245−13258.

(36) Peng, Q.; Duarte, F.; Paton, R. S. Computing organic stereoselectivity-from concepts to quantitative calculations and predictions. *Chem. Soc. Rev.* **2016**, *45*, 6093−6107.

(37) Sterling, A. J.; Zavitsanou, S.; Ford, J.; Duarte, F. Selectivity in organocatalysis—From qualitative to quantitative predictive models. *Wiley Interdiscip. Rev.:Comput. Mol. Sci.* **2021**, *11*, No. e1518.

(38) Young, T. A.; Silcock, J. J.; Sterling, A. J.; Duarte, F. autodE: automated calculation of reaction energy profiles—application to organic and organometallic reactions. *Angew. Chem.* **2021**, *133*, 4312−4320.

(39) Balcells, D.; Clot, E.; Eisenstein, O.; Nova, A.; Perrin, L. Deciphering selectivity in organic reactions: a multifaceted problem. *Acc. Chem. Res.* **2016**, *49*, 1070−1078.

(40) Wang, S.; Witek, J.; Landrum, G. A.; Riniker, S. Improving conformer generation for small rings and macrocycles based on distance geometry and experimental torsional-angle preferences. *J. Chem. Inf. Model.* **2020**, *60*, 2044−2058.

https://doi.org/10.1021/acs.jcim.5c02074
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

(41) Hawkins, P. C. D.; Skillman, A. G.; Nicholls, A. Comparison of shape-matching and docking as virtual screening tools. *J. Med. Chem.* **2007**, *50*, 74−82.

(42) Gómez-Bombarelli, R.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Duvenaud, D.; Maclaurin, D.; Blood-Forsythe, M. A.; Chae, H. S.; Einzinger, M.; Ha, D.-G.; Wu, T.; et al. Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach. *Nat. Mater.* **2016**, *15*, 1120−1127.

(43) Shu, Y.; Levine, B. G. Simulated evolution of fluorophores for light emitting diodes. *J. Chem. Phys.* **2015**, *142*, No. 104104, DOI: 10.1063/1.4914294.

(44) Lin, K.-H.; Wetzelaer, G.-J. A.; Blom, P. W.; Andrienko, D. Virtual Screening of TADF Emitters for Single-Layer OLEDs. *Front. Chem.* **2021**, *9*, No. 800027.

(45) Cheng, C. Y.; Campbell, J. E.; Day, G. M. Evolutionary chemical space exploration for functional materials: computational organic semiconductor discovery. *Chem. Sci.* **2020**, *11*, 4922−4933.

(46) Balcells, D.; Skjelstad, B. B. tmQM dataset—quantum geometries and properties of 86k transition metal complexes. *J. Chem. Inf. Model.* **2020**, *60*, 6135−6146.

(47) Jin, H.; Merz, K. M., Jr Modeling Zinc Complexes Using Neural Networks. *J. Chem. Inf. Model.* **2024**, *64*, 3140−3148.

(48) 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*, 7642−7654.

(49) Kevlishvili, I.; Duan, C.; Kulik, H. J. Classification of hemilabile ligands using machine learning. *J. Phys. Chem. Lett.* **2023**, *14*, 11100−11109.

(50) Nandy, A.; Taylor, M. G.; Kulik, H. J. Identifying underexplored and untapped regions in the chemical space of transition metal complexes. *J. Phys. Chem. Lett.* **2023**, *14*, 5798−5804.

(51) Nandy, A.; Duan, C.; Taylor, M. G.; Liu, F.; Steeves, A. H.; Kulik, H. J. Computational discovery of transition-metal complexes: from high-throughput screening to machine learning. *Chem. Rev.* **2021**, *121*, 9927−10000.

(52) Kneiding, H.; Lukin, R.; Lang, L.; Reine, S.; Pedersen, T. B.; De Bin, R.; Balcells, D. Deep learning metal complex properties with natural quantum graphs. *Digital Discovery* **2023**, *2*, 618−633.

(53) Vela, S.; Laplaza, R.; Cho, Y.; Corminboeuf, C. cell2mol: encoding chemistry to interpret crystallographic data. *npj Comput. Mater.* **2022**, *8*, No. 188.

(54) Arunachalam, N.; Gugler, S.; Taylor, M. G.; Duan, C.; Nandy, A.; Janet, J. P.; Meyer, R.; Oldenstaedt, J.; Chu, D. B.; Kulik, H. J. Ligand additivity relationships enable efficient exploration of transition metal chemical space. *J. Chem. Phys.* **2022**, *157*, No. 184112, DOI: 10.1063/5.0125700.

(55) Piskorz, T. K.; Marti-Centelles, V.; Young, T. A.; Lusby, P. J.; Duarte, F. Computational Modeling of Supramolecular Metallo-organic Cages-Challenges and Opportunities. *ACS Catal.* **2022**, *12*, 5806−5826.

(56) Cho, Y.; Laplaza, R.; Vela, S.; Corminboeuf, C. Automated prediction of ground state spin for transition metal complexes. *Digital Discovery* **2024**, *3*, 1638−1647.

(57) 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*, 2294−2308.

(58) Ioannidis, E. I.; Gani, T. Z. H.; Kulik, H. J. molSimplify: A toolkit for automating discovery in inorganic chemistry. *J. Comput. Chem.* **2016**, *37*, 2106−2117.

(59) Sobez, J.-G.; Reiher, M. Molassembler: Molecular Graph Construction, Modification, and Conformer Generation for Inorganic and Organic Molecules. *J. Chem. Inf. Model.* **2020**, *60*, 3884−3900.

(60) Taylor, M. G.; Burrill, D. J.; Janssen, J.; Batista, E. R.; Perez, D.; Yang, P. Architector for high-throughput cross-periodic table 3D complex building. *Nat. Commun.* **2023**, *14*, No. 2786.

(61) Chernyshov, I. Y.; Pidko, E. A. MACE: Automated Assessment of Stereochemistry of Transition Metal Complexes and Its Applications in Computational Catalysis. *J. Chem. Theory Comput.* **2024**, *20*, 2313−2320.

(62) Foscato, M.; Venkatraman, V.; Jensen, V. R. DENOPTIM: Software for Computational de Novo Design of Organic and Inorganic Molecules. *J. Chem. Inf. Model.* **2019**, *59*, 4077−4082.

(63) Turcani, L.; Tarzia, A.; Szczypiński, F. T.; Jelfs, K. E. stk: An extendable Python framework for automated molecular and supramolecular structure assembly and discovery. *J. Chem. Phys.* **2021**, *154*, No. 214102.

(64) Clarke, C.; Sommer, T.; Kleuker, F.; García-Melchor, M. *DART: Unlocking Coordination Chemistry Beyond the Cambridge Structural Database.* 2024; https://chemrxiv.org/engage/chemrxiv/article-details/6717eb4e83f22e4214d2b98b.

(65) Young, T. A.; Gheorghe, R.; Duarte, F. cgbind: A python module and web app for automated metallocage construction and host-guest characterization. *J. Chem. Inf. Model.* **2020**, *60*, 3546−3557.

(66) Jin, H.; Merz, K. M., Jr LigandDiff: de novo ligand design for 3D transition metal complexes with diffusion models. *J. Chem. Theory Comput.* **2024**, *20*, 4377−4384.

(67) Jin, H.; Merz, K. M., Jr Partial to total generation of 3D transition-metal complexes. *J. Chem. Theory Comput.* **2024**, *20*, 8367−8377.

(68) Cornet, F.; Benediktsson, B.; Hastrup, B.; Schmidt, M. N.; Bhowmik, A. OM-Diff: inverse-design of organometallic catalysts with guided equivariant denoising diffusion. *Digital Discovery* **2024**, *3*, 1793−1811.

(69) Cornet, F. R. J.; Deshmukh, P.; Benediktsson, B.; Schmidt, M. N.; Bhowmik, A. Equivariant conditional diffusion model for exploring the chemical space around Vaska's complex *AI for Accelerated Materials Design - NeurIPS 2024* 2024.

(70) Cornet, F.; Deshmukh, P.; Benediktsson, B.; Schmidt, M. N.; Bhowmik, A. Improving generative inverse design of molecular catalysts in small data regime. *Mach. Learn. Sci. Technol.* **2025**, *6*, No. 025057.

(71) Laplaza, R.; Sobez, J.-G.; Wodrich, M. D.; Reiher, M.; Corminboeuf, C. The (not so) simple prediction of enantioselectivity-a pipeline for high-fidelity computations. *Chem. Sci.* **2022**, *13*, 6858−6864.

(72) Nandy, A.; Duan, C.; Goffinet, C.; Kulik, H. J. New strategies for direct methane-to-methanol conversion from active learning exploration of 16 million catalysts. *Jacs Au* **2022**, *2*, 1200−1213.

(73) Adamji, H.; Nandy, A.; Kevlishvili, I.; Román-Leshkov, Y.; Kulik, H. J. Computational discovery of stable metal-organic frameworks for methane-to-methanol catalysis. *J. Am. Chem. Soc.* **2023**, *145*, 14365−14378.

(74) Friederich, P.; dos Passos Gomes, G.; De Bin, R.; Aspuru-Guzik, A.; Balcells, D. Machine learning dihydrogen activation in the chemical space surrounding Vaska's complex. *Chem. Sci.* **2020**, *11*, 4584−4601.

(75) Ye, Y. S.; Laverny, A.; Wodrich, M. D.; Laplaza, R.; Fadaei-Tirani, F.; Scopelliti, R.; Corminboeuf, C.; Cramer, N. Enantiospecific Synthesis of Planar Chiral Rhodium and Iridium Cyclopentadienyl Complexes: Enabling Streamlined and Computer-Guided Access to Highly Selective Catalysts for Asymmetric C-H Functionalizations. *J. Am. Chem. Soc.* **2024**, *146*, 34786−34795.

(76) Morán-González, L.; Burnage, A. L.; Nova, A.; Balcells, D. AI Approaches to Homogeneous Catalysis with Transition Metal Complexes. *ACS Catal.* **2025**, *15*, 9089−9105.

(77) Terrones, G. G.; Huang, S.-P.; Rivera, M. P.; Yue, S.; Hernandez, A.; Kulik, H. J. Metal-organic framework stability in water and harsh environments from data-driven models trained on the diverse WS24 data set. *J. Am. Chem. Soc.* **2024**, *146*, 20333−20348.

(78) Nandy, A.; Terrones, G.; Arunachalam, N.; Duan, C.; Kastner, D. W.; Kulik, H. J. MOFSimplify, machine learning models with extracted stability data of three thousand metal−organic frameworks. *Sci. Data* **2022**, *9*, No. 74.

https://doi.org/10.1021/acs.jcim.5c02074
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

(79) Mariano, L. A.; Nguyen, V. H. A.; Briganti, V.; Lunghi, A. Charting Regions of Cobalt's Chemical Space with Maximally Large Magnetic Anisotropy: A Computational High-Throughput Study. *J. Am. Chem. Soc.* **2024**, *146*, 34158−34166.

(80) Huang, X.; Kevlishvili, I.; Craig, S. L.; Kulik, H. J. Force-Activated Spin-Crossover in Fe2+ and Co2+ Transition Metal Mechanophores. *Inorg. Chem.* **2025**, *64*, 380−392.

(81) Janet, J. P.; Chan, L.; Kulik, H. J. Accelerating chemical discovery with machine learning: simulated evolution of spin crossover complexes with an artificial neural network. *J. Phys. Chem. Lett.* **2018**, *9*, 1064−1071.

(82) Frangoulis, L.; Khatibi, Z.; Mariano, L. A.; Lunghi, A. Generating New Coordination Compounds via Multireference Simulations, Genetic Algorithms, and Machine Learning: The Case of Co (II) and Dy (III) Molecular Magnets. *JACS Au* **2025**, *5*, 3808−3821.

(83) Vennelakanti, V.; Jeon, M.; Kulik, H. J. How Do Differences in Electronic Structure Affect the Use of Vanadium Intermediates as Mimics in Nonheme Iron Hydroxylases? *Inorg. Chem.* **2024**, *63*, 4997−5011.

(84) Terrones, G. G.; Duan, C.; Nandy, A.; Kulik, H. J. Low-cost machine learning prediction of excited state properties of iridium-centered phosphors. *Chem. Sci.* **2023**, *14*, 1419−1433.

(85) Duan, C.; Nandy, A.; Terrones, G. G.; Kastner, D. W.; Kulik, H. J. Active learning exploration of transition-metal complexes to discover method-insensitive and synthetically accessible chromophores. *Jacs Au* **2023**, *3*, 391−401.

(86) Summers, T. J.; Taylor, M. G.; Augustine, L. J.; Janssen, J.; Perez, D.; Batista, E. R.; Yang, P. On the importance of configuration search to the predictivity of lanthanide selectivity. *JACS Au* **2024**, *5* (2), 631−641, DOI: 10.1021/jacsau.4c00770.

(87) Pita-Milleiro, A.; Strandgaard, M.; Linjordet, T.; Kneiding, H.; Balcells, D. Evolving Light Harvesting Metal Complexes with AI-Made Ligands *ChemRxiv* 2025.

(88) Jin, H.; Merz, K. M., Jr Toward AI/ML-assisted discovery of transition metal complexes. *Annu. Rep. Comput. Chem.* **2024**, *20*, 225−267.

(89) Toney, J. W.; Garrison, A. G.; Luo, W.; Michel, R. G. S.; Mukhopadhyay, S.; Kulik, H. J. Exploring beyond experiment: generating high-quality datasets of transition metal complexes with quantum chemistry and machine learning. *Curr. Opin. Chem. Eng.* **2025**, *50*, No. 101189.

(90) Cowan, A. J.; George, M. W. Formation and reactivity of organometallic alkane complexes. *Coord. Chem. Rev.* **2008**, *252*, 2504−2511.

(91) Kubas, G. J. Molecular hydrogen complexes: coordination of a. sigma. bond to transition metals. *Acc. Chem. Res.* **1988**, *21*, 120−128.

(92) Valente, C.; Çalimsiz, S.; Hoi, K. H.; Mallik, D.; Sayah, M.; Organ, M. G. The development of bulky palladium NHC complexes for the most-challenging cross-coupling reactions. *Angew. Chem., Int. Ed.* **2012**, *51*, 3314−3332.

(93) Vougioukalakis, G. C.; Grubbs, R. H. Ruthenium-based heterocyclic carbene-coordinated olefin metathesis catalysts. *Chem. Rev.* **2010**, *110*, 1746−1787.

(94) Hagen, H.; Boersma, J.; van Koten, G. Homogeneous vanadium-based catalysts for the Ziegler-Natta polymerization of α-olefins. *Chem. Soc. Rev.* **2002**, *31*, 357−364.

(95) Mills, L. R.; Edjoc, R. K.; Rousseaux, S. A. Design of an electron-withdrawing benzonitrile ligand for Ni-catalyzed cross-coupling involving tertiary nucleophiles. *J. Am. Chem. Soc.* **2021**, *143*, 10422−10428.

(96) Chen, S.; Nielson, T.; Zalit, E.; Skjelstad, B. B.; Borough, B.; Hirschi, W. J.; Yu, S.; Balcells, D.; Ess, D. H. Automated construction and optimization combined with machine learning to generate Pt (II) methane C-H activation transition states. *Top. Catal.* **2022**, *65*, 312−324.

(97) Dohm, S.; Hansen, A.; Steinmetz, M.; Grimme, S.; Checinski, M. P. Comprehensive Thermochemical Benchmark Set of Realistic Closed-Shell Metal Organic Reactions. *J. Chem. Theory Comput.* **2018**, *14*, 2596−2608.

(98) 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. Cheminf.* **2025**, *17*, No. 63.

(99) Groom, C. R.; Bruno, I. J.; Lightfoot, M. P.; Ward, S. C. The Cambridge structural database. *Struct. Sci.* **2016**, *72*, 171−179.

(100) Adamo, C.; Barone, V. Toward reliable density functional methods without adjustable parameters: The PBE0 model. *J. Chem. Phys.* **1999**, *110*, 6158−6170.

(101) Grimme, S.; Ehrlich, S.; Goerigk, L. Effect of the damping function in dispersion corrected density functional theory. *J. Comput. Chem.* **2011**, *32*, 1456−1465.

(102) Weigend, F.; Ahlrichs, R. Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3297.

(103) Pattanaik, L.; Ingraham, J. B.; Grambow, C. A.; Green, W. H. Generating transition states of isomerization reactions with deep learning. *Phys. Chem. Chem. Phys.* **2020**, *22*, 23618−23626.

(104) Choi, S. Prediction of transition state structures of gas-phase chemical reactions via machine learning. *Nat. Commun.* **2023**, *14*, No. 1168.

(105) Kim, S.; Woo, J.; Kim, W. Y. Diffusion-based generative AI for exploring transition states from 2D molecular graphs. *Nat. Commun.* **2024**, *15*, No. 341.

(106) Pritchard, B. P.; Altarawy, D.; Didier, B.; Gibsom, T. D.; Windus, T. L. A New Basis Set Exchange: An Open, Up-to-date Resource for the Molecular Sciences Community. *J. Chem. Inf. Model.* **2019**, *59*, 4814−4820.

(107) Frisch, M. J. et al. *Gaussian 16. Revision B.01*. 2016; Gaussian Inc: Wallingford CT.

(108) Pracht, P.; Bohle, F.; Grimme, S. Automated exploration of the low-energy chemical space with fast quantum chemical methods. *Phys. Chem. Chem. Phys.* **2020**, *22*, 7169−7192.

(109) Park, S. H.; Kwak, J.; Shin, K.; Ryu, J.; Park, Y.; Chang, S. Mechanistic Studies of the Rhodium-Catalyzed Direct C-H Amination Reaction Using Azides as the Nitrogen Source. *J. Am. Chem. Soc.* **2014**, *136*, 2492−2502.

(110) Kim, S.-T.; Strauss, M. J.; Cabré, A.; Buchwald, S. L. Room-Temperature Cu-Catalyzed Amination of Aryl Bromides Enabled by DFT-Guided Ligand Design. *J. Am. Chem. Soc.* **2023**, *145*, 6966−6975.

(111) Haibach, M. C.; Guan, C.; Wang, D. Y.; Li, B.; Lease, N.; Steffens, A. M.; Krogh-Jespersen, K.; Goldman, A. S. Olefin Hydroaryloxylation Catalyzed by Pincer-Iridium Complexes. *J. Am. Chem. Soc.* **2013**, *135*, 15062−15070.

(112) Singh, D.; Buratto, W. R.; Torres, J. F.; Murray, L. J. Activation of dinitrogen by polynuclear metal complexes. *Chem. Rev.* **2020**, *120*, 5517−5581.

(113) Jasniewski, A. J.; Que, L., Jr Dioxygen activation by nonheme diiron enzymes: diverse dioxygen adducts, high-valent intermediates, and related model complexes. *Chem. Rev.* **2018**, *118*, 2554−2592.

(114) Dashtian, K.; Shahsavarifar, S.; Usman, M.; Joseph, Y.; Ganjali, M. R.; Yin, Z.; Rahimi-Nasrabadi, M. A comprehensive review on advances in polyoxometalate based materials for electrochemical water splitting. *Coord. Chem. Rev.* **2024**, *504*, No. 215644.

(115) Lee, K.; Park, S.; Park, M. Supporting Information for MetalloGen: Automated 3D Conformer Generation for Diverse Coordination Complexes. 2025 DOI: 10.6084/m9.figshare.29999380.