5 Supplementary Material

5.1 Evaluation details

5.1.1 Validity, Uniqueness, Novelty

All the presented metrics are proportions of the generated samples.

Validity A generated complex has to pass a series of checks to be deemed valid:

1. **(one TM check)** It has to have **exactly** one transition metal atom;
2. **(distance check)** All pairwise distances should be at least 0.9Å, and no atom can be disconnected from the rest of the complex (i.e. its closest neighbour is located at distance larger than the cutoff of 3.0Å);
3. **(RDKit check)** The ligands, i.e. complex where the TM has been removed, have to be valid according to RDKit [25].

As the algorithm implemented in RDKit to determine bonds can not handle transition metals, we proceed as follows: we remove the metal centre, and we then try to find a feasible bond allocation using `rdDetermineBonds.DetermineBonds`. If the allocation succeeds, the sample is deemed valid. As the removal of the metal centre can introduce local charges, we apply `rdDetermineBonds.DetermineBonds` for different charges until one matches. If none matches, the configuration is deemed invalid.

The validation method is not perfect, as only around 88% of the training database is deemed valid by our algorithm (Table 1).

Uniqueness and Novelty As bonding is not properly defined for transition metal complexes, we study uniqueness and novelty in terms of chemical formulas. This does not provide the full picture as two identical formulas can correspond to different complexes. However, when encountering new formulas, we are ensured that the corresponding complexes are novel. Uniqueness and Novelty are defined as follows,

$$\text{UF} = \frac{\text{\# (valid and unique formulas)}}{\text{\# samples}}, \tag{7}$$

$$\text{NF} = \frac{\text{\# (valid, unique and novel formulas)}}{\text{\# samples}}. \tag{8}$$

5.1.2 Geometry and Binding Energy

Given the importance of the direct neighbourhood of the centre, we assess the geometry of centre and the two proximal atoms by comparing the empirical distribution of the $\text{L}_{1,2} - \text{M}$ distances and the $\text{L}_1 - \text{M} - \text{L}_2$ angle. Similarly, we compare the training distribution of binding energy with the distribution induced by the generated samples.

We measure the discrepancy between training distributions and distributions induced by the generated samples using the 1-Wassertein distance. If P_z denotes the empirical measure for centre $z \in \mathcal{Z}$ across the dataset, and Q_z denotes the empirical measure the same centre across the samples generated by the diffusion model, the distance between the two empirical distributions is given by

$$W(P_z, Q_z) = \big(\frac{1}{n}\sum_{i=1}^{n} ||X_{(i)} - Y_{(i)}||\big), \tag{9}$$

where $X_{(i)}$ and $Y_{(i)}$ denote samples from P_z and Q_z respectively.

To obtain an aggregated distance value, we compute a weighted sum over the different metal-centres,

$$W(P, Q) = \sum_{z \in \mathcal{Z}} p(z) W(P_z, Q_z), \tag{10}$$

where $p(z)$ denotes the empirical categorical distribution over the metal centre obtained from the training data.

5.1.3 Baselines

Our method, coined **OM-EDM-PAINN** implements Eqs. (1) and (3), and the more expressive denoising neural network inspired from the PAINN architecture [17]. We compare it with 3 different baselines:

- **EDM**: that reimplements the vanilla equivariant denoising diffusion [9];
- **EDM-PAINN**: that reimplements the vanilla equivariant denoising diffusion with a more expressive denoising neural network identical to that of **OM-EDM-PAINN**;
- **OM-EDM**: that implements Eqs. (1) and (3), but uses EGNN [18] as denosing neural network.

We additionally include a baseline based on geometry-free representations, where RDKit, or any cheap force-field, is used to generate geometries. We denote that method by FF. We compute the geometry statistics from the additional 18064 force-field-level data points released along with the DFT-level data [22].