OM-DIFF: INVERSE-DESIGN OF ORGANOMETALLIC CATALYSTS WITH GUIDED EQUIVARIANT DENOISING DIFFUSION

SUPPLEMENTARY MATERIAL

François Cornet
Department of Applied Mathematics and Computer Science
Technical University of Denmark
Kgs. Lyngby 2800, Denmark
frjc@dtu.dk

Bardi Benediktsson
Department of Energy Conversion and Storage
Technical University of Denmark
Kgs. Lyngby 2800, Denmark

Bjarke Hastrup
Department of Energy Conversion and Storage
Technical University of Denmark
Kgs. Lyngby 2800, Denmark

Mikkel N. Schmidt
Department of Applied Mathematics and Computer Science
Technical University of Denmark
Kgs. Lyngby 2800, Denmark

Arghya Bhowmik*
Department of Energy Conversion and Storage
Technical University of Denmark
Kgs. Lyngby 2800, Denmark
arbh@dtu.dk

S1 Implementation Details

S1.1 Data splits

In this section, we detail how the data was split.

Diffusion model and time-conditioned regressor We use the same split to train the diffusion model and time-conditioned regressor. We partition the data at random as follows: 90% for training (6348 data points), 5% for validation (352 data points), and 5% for testing (354 data points). We deviate from the usual $80-10-10$ split, as the performance of the generative model is evaluated on generated samples, and not on a held-out fold. We therefore made the validation and test folds as small as possible, while keeping a sufficient amount of data points to allow for meaningful estimation of the error metrics (e.g. for early stopping of the regressor model). Figs. 7b and S12 are obtained by evaluating the error of the time-conditioned regressor on the test fold. The purpose of that evaluation is qualitative rather than quantitative. The same holds for Fig. S1, where the different diffusion models are evaluated on the test fold.

Screening regressor To obtain the numbers reported in Table 2, we perform a stratified 10-fold cross-validation, where we make sure that (1) the proportion of metal centers is kept constant across folds, (2) the property range is covered uniformly. For each test fold, the screening surrogate is consequently trained on the remaining 90% of the data (further split in $80-10$ for train-validation). The standard deviation is reported across folds.

To perform the production run, we retrain one model on 95% of the data (6701 data points), and keep 5% (353 data points) for early stopping.

S1.2 Denoiser architecture

In this section, we provide additional details regarding the architecture of the different variants of the denoising neural network ε_θ. As a recall, ε_θ maps a noisy atomistic structure $\mathcal{C}_t$ and a time step t to a noise estimate $[\hat{\epsilon}_{x^{(L)}}, \hat{\epsilon}_{h^{(L)}}]$.

OM-DIFF$(\checkmark, \{\checkmark, \mathbf{x}\})$ These variants implement an architecture similar to that of EDM[48], based on EGNN[78].

OM-DIFF$(\{\checkmark, \mathbf{x}\}, \checkmark)$ These variants implement an improved architecture ε_θ. Internally, each atom i is given a hidden state defined by a tuple $\mathbf{h}_i^m = (\mathbf{s}_i^m, \mathbf{v}_i^m)$ where $\mathbf{s}_i^m \in \mathbb{R}^{1\times D}$ is a scalar feature vector, and $\mathbf{v}_i^m \in \mathbb{R}^{3\times D}$ a set of vectorial features. Initially, $\mathbf{s}_i^0$ is obtained via a linear projection of the one-hot encoded atom type, while $\mathbf{v}_i^0$ is set to $\mathbf{0}$. Time is featurized through 16 random Fourier features, and concatenated with each atom scalar features. Then, $\mathbf{h}_i^0$ gets updated through M successive message-passing rounds. We employ message and update blocks similar to those of PAINN[79]. Connectivity is defined with a 7.5Å cutoff, and for each edge we keep track of a scalar state, that also gets updated at each message passing step through a simple one layer MLP that maps the current edge state and the states of the two corresponding atoms to the new edge state. The initial edge states are obtained by featurizing pairwise distances through Gaussian radial basis functions. After the message-passing phase, the final states $\mathbf{h}_i^M$ are read out to produce $[\hat{\epsilon}_{x^{(L)}}, \hat{\epsilon}_{h^{(L)}}]$. A gated equivariant block[79] is employed to obtain $\hat{\epsilon}_{x^{(L)}}$ from $\mathbf{h}_i^M$, while $\mathbf{s}_i^M$ is processed through a one hidden-layer MLP to obtain $\hat{\epsilon}_{h^{(L)}}$. The different hidden sizes are kept constant throughout the network. The most important hyperparameters are summarized in Table S1.

Table S1: Hyper-parameters setup for the denoiser architecture ε_θ

Hyper-parameter	**Value**
Number of interactions size	5
Hidden node size (D)	256
Edge size	64
Activation functions	SILU
RBF	Gaussian
Cutoff	7.5 Å
Optimizer	AdamW
Learning rate	10^{-4}
Weight decay	10^{-12}
Denoising steps (T)	1000
Noise schedule	VP-Polynomial[48]

S1.3 Denoiser performance

In Fig. S1, we display the loss of the different variants of OM-DIFF. The loss is evaluated on a held-out test fold. The left column displays the errors related to h, i.e. the atom types, while the right column displays the errors related to x, i.e. the coordinates. The top row shows the noise prediction error – similar to the training objective of the neural network. The bottom row displays the resulting error on the estimated denoised sample obtained using the relationship $\mathcal{C}^{(L)} = \frac{1}{\alpha_t}(\mathcal{C}_t^{(L)} - \sigma_t \epsilon)$.

S2 Evaluation of sampled complexes

In this section, we provide additional details regarding the evaluation of the complexes sampled from our generative model.

S2.1 Validity, Uniqueness, Novelty

All the reported numbers are expressed as 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 atoms should have the distance to their nearest neighbour that falls within a specified range,

$$\forall i : \min_{j\neq i} d_{ij} \in [d_{z_i}^{\min}; d_{z_i}^{\max}] \tag{S1}$$

where $d_{z_i}^{\min}$ and $d_{z_i}^{\max}$ are minimal and maximal distances to nearest neighbour for atom of type z_i across the training database (99% percentile $\pm$ a 10 % margin);

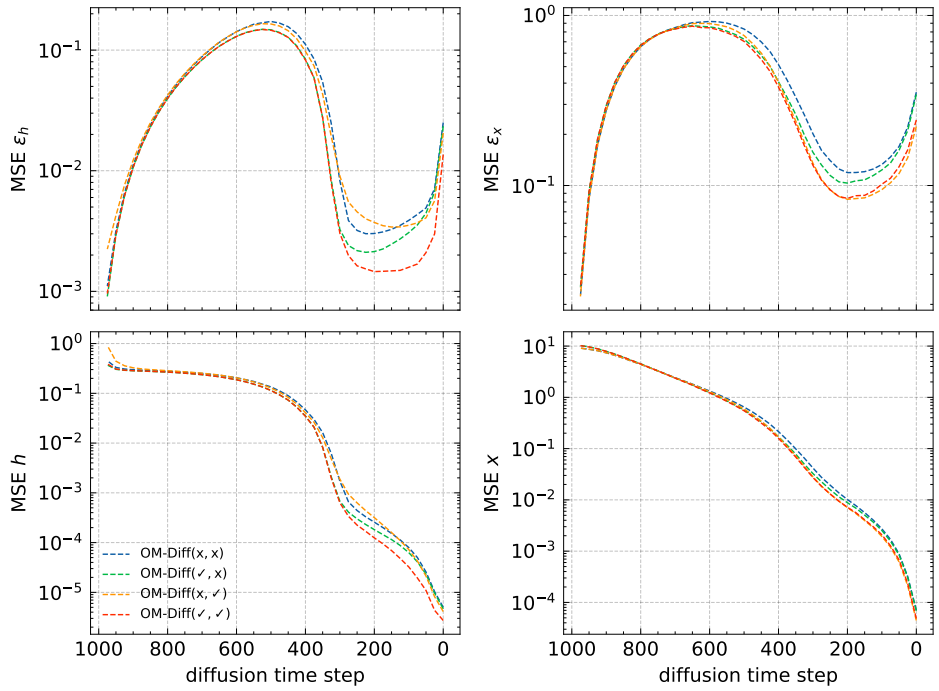


Figure S1: Evaluation of the different denoisers. **(Top row)** Prediction error for the different denoisers. **(Bottom row)** Resulting error on the estimated denoised sample for each denoiser.

3. **(RDKit check)** The ligands, i.e. complex where the TM has been removed, have to be valid according to RDKit[51]. As the algorithm implemented in RDKit to determine bonds can not handle transition metals, we proceed as follows: we remove the metal center, and we then use `rdDetermineBonds.DetermineBonds` (with `useHueckel=True`) on the remaining atoms. We do not allow charges as the training ligands are neutral. A sample is deemed valid if the bond allocation succeeds, and the inferred `Mol` object is composed of two distinct fragments, i.e. corresponding to L_1 and L_2.

While not bulletproof, the validation method classifies around 99% of the training database as valid. We provide the detailed validity results for unconditional sampling in Table S2.

Table S2: Detailed validity results for unconditional sampling. All presented numbers are expressed in % of the number of sampled complexes. Higher is better.

		OM-DIFF(x, x)	OM-DIFF(✓, x)	OM-DIFF(x, ✓)	OM-DIFF(✓, ✓)
Exactly one MC		99.28	100.00	96.57	100.00
Distance check	All	15.99	27.94	36.77	46.60
	Ni	16.27	28.67	37.59	46.37
	Cu	14.65	25.68	39.34	45.15
	Pd	16.35	27.27	38.33	47.93
	Ag	12.38	21.17	34.60	41.95
	Pt	20.39	34.63	40.29	50.56
	Au	18.50	31.55	37.21	46.06
RDKit check	All	8.95	19.57	18.88	28.22
	Ni	6.86	18.89	14.79	25.30
	Cu	7.12	15.10	17.06	23.50
	Pd	10.18	19.40	20.98	28.34
	Ag	6.62	14.51	16.84	25.25
	Pt	14.58	27.84	23.60	35.41
	Au	11.09	23.11	19.84	31.55

Uniqueness and Novelty Once a complex is deemed valid, we convert it to a tuple $\big(\mathtt{M}, \{\mathtt{L}_1, \mathtt{L}_2\}\big)$, where $\{\}$ denotes a multiset data-structure, i.e. unordered collection of elements which may be repeated. Uniqueness is defined as the ratio of unique tuples among all generated tuples. Novelty is defined as the ratio of unique and novel tuples, i.e. that are not part of the training database, among all generated tuples:

$$\text{V} = \frac{\text{\# valid}}{\text{\# samples}}, \tag{S2}$$

$$\text{V\&U} = \frac{\text{\# (valid and unique)}}{\text{\# samples}}, \tag{S3}$$

$$\text{V\&U\&N} = \frac{\text{\# (valid, unique and novel)}}{\text{\# samples}}. \tag{S4}$$

In Fig. S2, we show how the different metrics evolve as the number of sampled complexes increases.

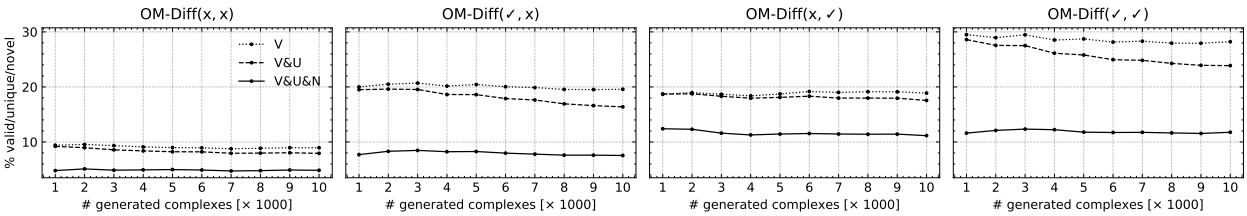


Figure S2: %Valid, %(Valid & Unique) and %(Valid & Unique & Novel) complexes for each variant of the generative model, as a function of the number of generated complexes. Novelty can come from novel combinations of existing compounds or from novel ligands.

Sources of novelty Due to the combinatorial nature of the training database, there are three possible sources of novelty:

$$\text{NC} = \text{the tuple } \big(\mathtt{M}, \{\mathtt{L}_1, \mathtt{L}_2\}\big) \text{ is novel, but } \mathtt{L}_1 \text{and } \mathtt{L}_2 \text{ are not,} \tag{S5}$$

$$\text{1L} = \text{either } \mathtt{L}_1 \text{or } \mathtt{L}_2 \text{ is novel,} \tag{S6}$$

$$\text{2L} = \text{both } \mathtt{L}_1 \text{and } \mathtt{L}_2 \text{ are novel.} \tag{S7}$$

In Fig. S3, how the number of novel ligands increases as the number of sampled complexes increases.

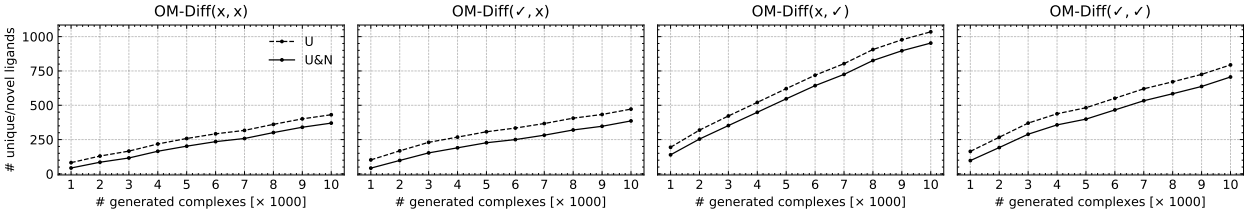


Figure S3: #Unique and #(Unique & Novel) ligands for each variant of the generative model, as a function of the number of generated complexes.

S2.2 Geometry and Binding Energy

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

$$W(P_z, Q_z) = \big(\frac{1}{n}\sum_{i=1}^{n} ||X_{(i)} - Y_{(i)}||\big), \tag{S8}$$

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-centers,

$$W(P, Q) = \sum_{z \in \mathcal{Z}} p(z) W(P_z, Q_z), \tag{S9}$$

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

Geometry around the metal center Given the importance of the direct neighbourhood of the center, we assess the geometry of central and the two proximal atoms by comparing the empirical distribution of the $L_{1,2} - M$ distances and the $L_1 - M - L_2$ angle. We provide all numerical results in Table S3, along with the corresponding distribution in Fig. S4, and the details of $L_{1,2}$-M for each metal center Fig. S5.

Table S3: Detailed geometry results for unconditional sampling. All presented numbers represent the Wassertein distance between the empirical histograms obtained from the dataset and generated samples. Lower is better.

	OM-DIFF(x, x)	OM-DIFF(✓, x)	OM-DIFF(x, ✓)	OM-DIFF(✓, ✓)	MMFF
$W^{L_{1,2}-M}$	0.2559	0.1647	0.2034	0.1468	0.8135
$W^{L_{1,2}-Ni}$	0.3215	0.2213	0.2279	0.1828	0.8164
$W^{L_{1,2}-Cu}$	0.2621	0.1355	0.1567	0.1129	0.9225
$W^{L_{1,2}-Pd}$	0.1680	0.1184	0.1389	0.0977	0.7440
$W^{L_{1,2}-Ag}$	0.3895	0.1907	0.2766	0.1678	0.8531
$W^{L_{1,2}-Pt}$	0.3365	0.2214	0.2527	0.2262	0.8681
$W^{L_{1,2}-Au}$	0.2903	0.2219	0.3001	0.2123	0.7960
$W^{L_1-M-L_2}$	0.0113	0.0084	0.0089	0.0079	0.0257
$W^{L_1-Ni-L_2}$	0.0139	0.0130	0.0119	0.0113	0.0330
$W^{L_1-Cu-L_2}$	0.0149	0.0092	0.0080	0.0065	0.0330
$W^{L_1-Pd-L_2}$	0.0073	0.0062	0.0066	0.0064	0.0259
$W^{L_1-Ag-L_2}$	0.0096	0.0089	0.0097	0.0102	0.0250
$W^{L_1-Pt-L_2}$	0.0153	0.0116	0.0132	0.0113	0.0198
$W^{L_1-Au-L_2}$	0.0136	0.0088	0.0108	0.0083	0.0194

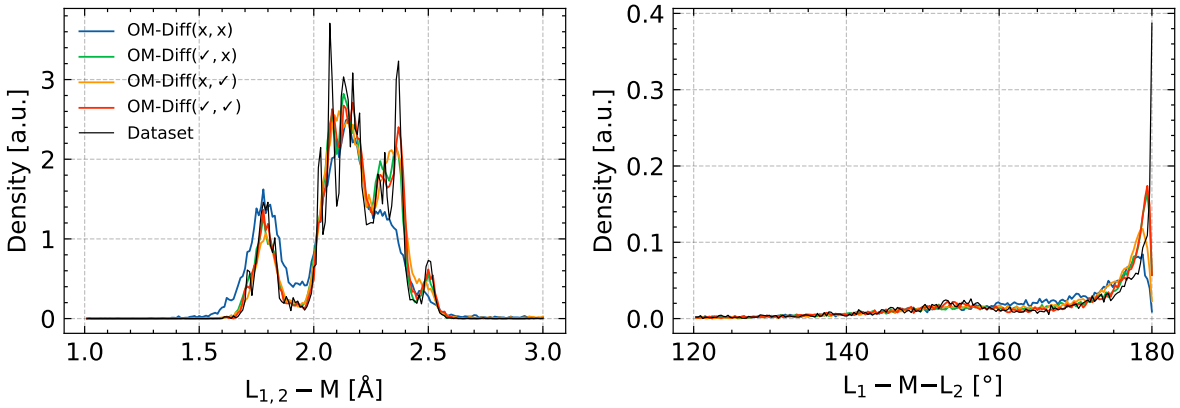


Figure S4: **(Top)** Distribution of $L_{1,2}$-M distances. **(Bottom)** Distribution of L_1-M-L_2 angles.

Binding energy Similarly, we compare the training distribution of binding energy with the distribution induced by the generated samples. The latter is estimated tas the mean prediction of an ensemble of 10 surrogate models. The numerical results are provided in Table S4 with the corresponding distributions being displayed in Fig. S6. In Fig. S7, we additionally provide the cumulative distribution of predictive uncertainty, estimated as the standard deviation across the same ensemble, detailed for each metal center.

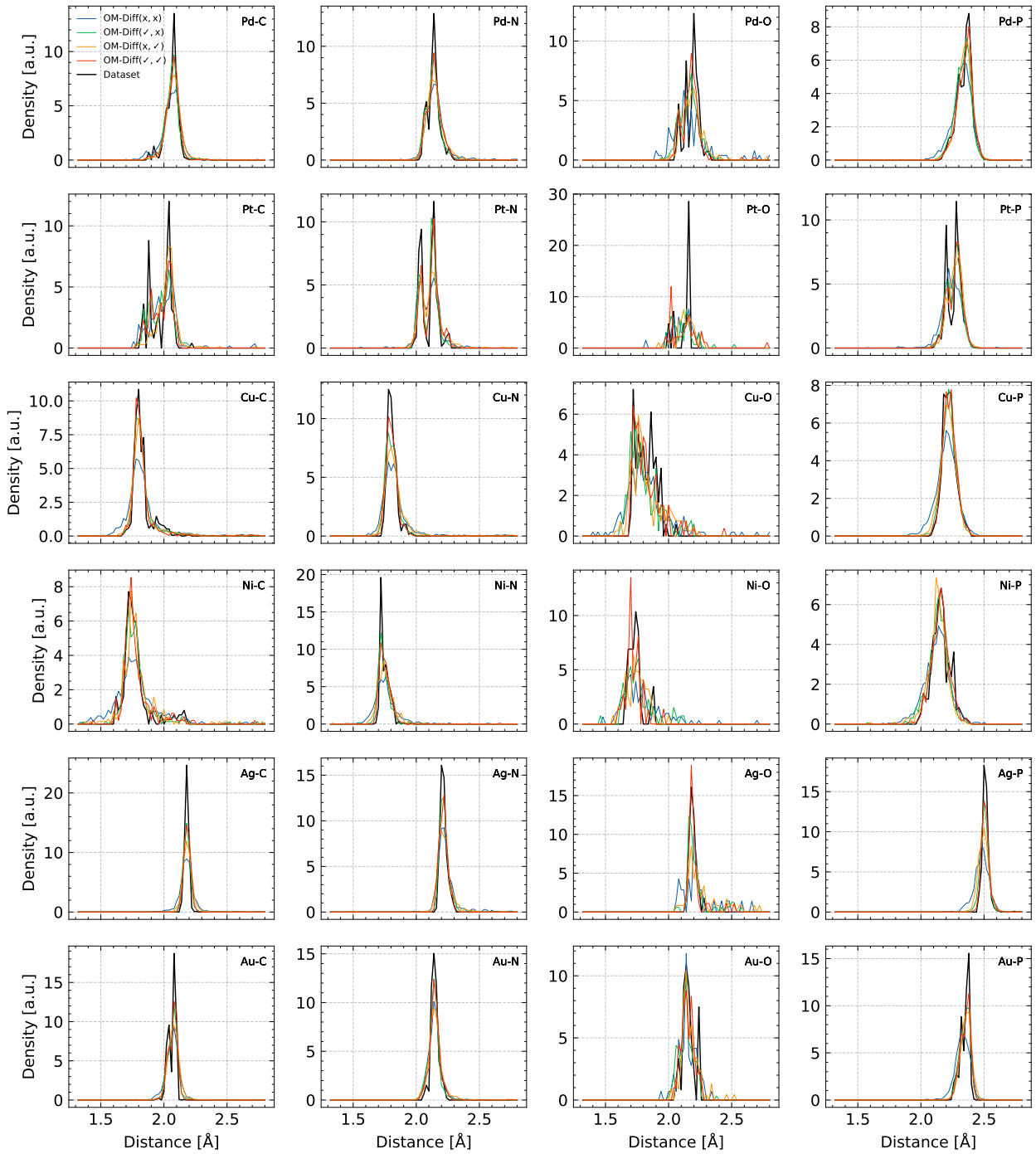


Figure S5: Distribution of bonds involving the metal center for the different variants of the diffusion generative model.

Table S4: Detailed energy results for unconditional sampling. All presented numbers represent the Wassertein distance between the empirical histograms obtained from the dataset and generated samples. Lower is better.

		OM-DIFF(x, x)	OM-DIFF(✓, x)	OM-DIFF(x, ✓)	OM-DIFF(✓, ✓)
$W^{\Delta E}$	All	0.0044	0.0040	0.0044	0.0036
	Ni	0.0071	0.0064	0.0061	0.0058
	Cu	0.0051	0.0049	0.0050	0.0043
	Pd	0.0032	0.0030	0.0028	0.0026
	Ag	0.0026	0.0024	0.0024	0.0025
	Pt	0.0085	0.0049	0.0059	0.0043
	Au	0.0040	0.0045	0.0065	0.0045

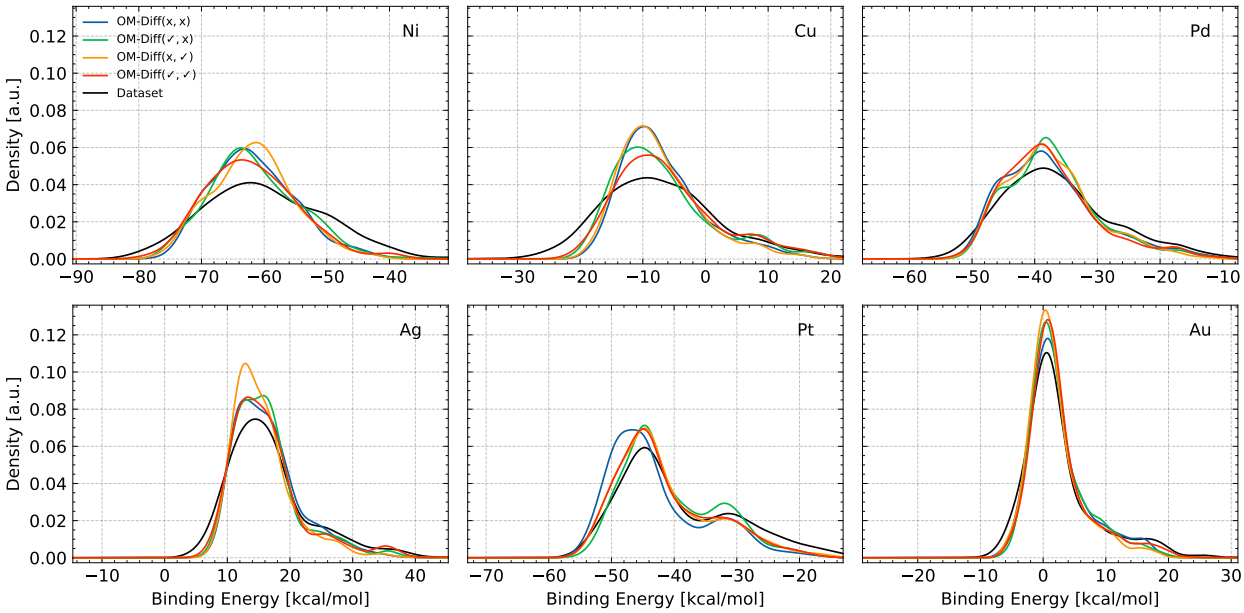


Figure S6: Distribution of binding energies for unconditionally generated samples.

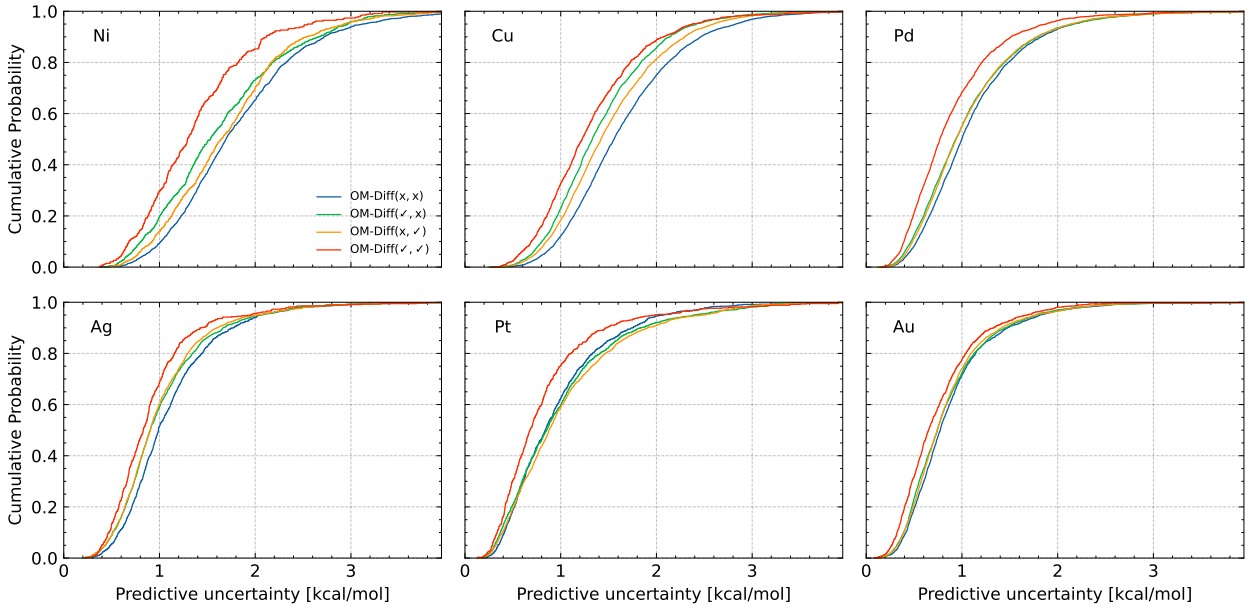


Figure S7: Predictive uncertainty, as estimated by an ensemble of 10 surrogates, evaluated on the samples generated by the different variants of the generative diffusion model. Similar to Fig. 6a, with each metal centre illustrated separately.

S2.3 Chemical composition

We compare the chemical composition of the sampled complexes and that of the training data, by comparing marginal distributions of atom types through total variation. We also compare distributions of molecular weights in Fig. S9, as molecular weight act as a proxy for the joint distribution of atom types.

Total variation We measure the discrepancy between two empirical categorical distributions as the total variation between the histogram obtained from generated samples, Q, and the histogram obtained from the training data, P. It writes

$$\mathrm{TV}(P, Q) = \sum_{z \in \mathcal{Z}} |P_z - Q_z|, \tag{S10}$$

where P_z refers to the average count for category z across the training data, and Q_z refers to the same quantity computed across the generated samples.

Total variation of metal-centers In this case, we compare the distributions of metal centers across generated samples and training data. For variants OM-DIFF$(\checkmark, \{\mathrm{x}, \checkmark\})$, the total variation is virtually 0, as they directly sample from the empirical distribution. We provide the metal center distribution in Table S5.

Table S5: Metal center distribution in % for unconditional sampling. The last column refers to the variants of OM-DIFF where the center is fixed.

	OM-DIFF(x, x)	OM-DIFF(x, ✓)	OM-DIFF(✓, {x, ✓})
Ni	12.97	9.45	5.93
Cu	28.25	25.67	18.34
Pd	27.21	32.19	37.26
Ag	9.21	8.29	10.06
Pt	7.75	9.02	8.98
Au	13.89	12.59	19.43

Total variation of non-TM elements We compute the total variation for all atom types except transition metals.

Total variation of proximal atoms We compute the total variation for proximal atoms, i.e. atoms that bind to the metal center. In Table S6, we provide the detailed numerical values for each metal center, and the corresponding distributions in Fig. S8.

Table S6: Detailed $\mathrm{TV}_{\mathrm{prox}}$, i.e. total variation of proximal atomic elements, in unconditional sampling. Lower is better.

	OM-DIFF(x, x)	OM-DIFF(✓, x)	OM-DIFF(x, ✓)	OM-DIFF(✓, ✓)
All	0.082	0.028	0.028	0.019
Ni	0.110	0.032	0.014	0.021
Cu	0.074	0.023	0.040	0.039
Pd	0.068	0.022	0.025	0.011
Ag	0.058	0.035	0.028	0.014
Pt	0.181	0.017	0.043	0.033
Au	0.088	0.050	0.013	0.009

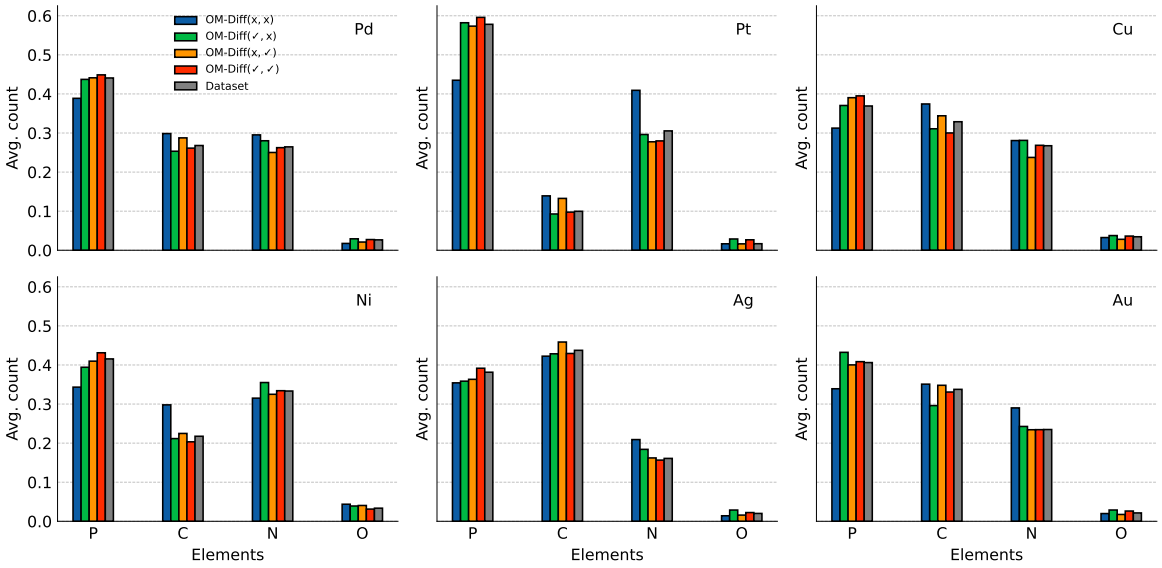


Figure S8: Distribution of proximal atoms, in unconditional sampling. Proximal atoms are defined as the atoms of the ligands to which the metal center is bound.

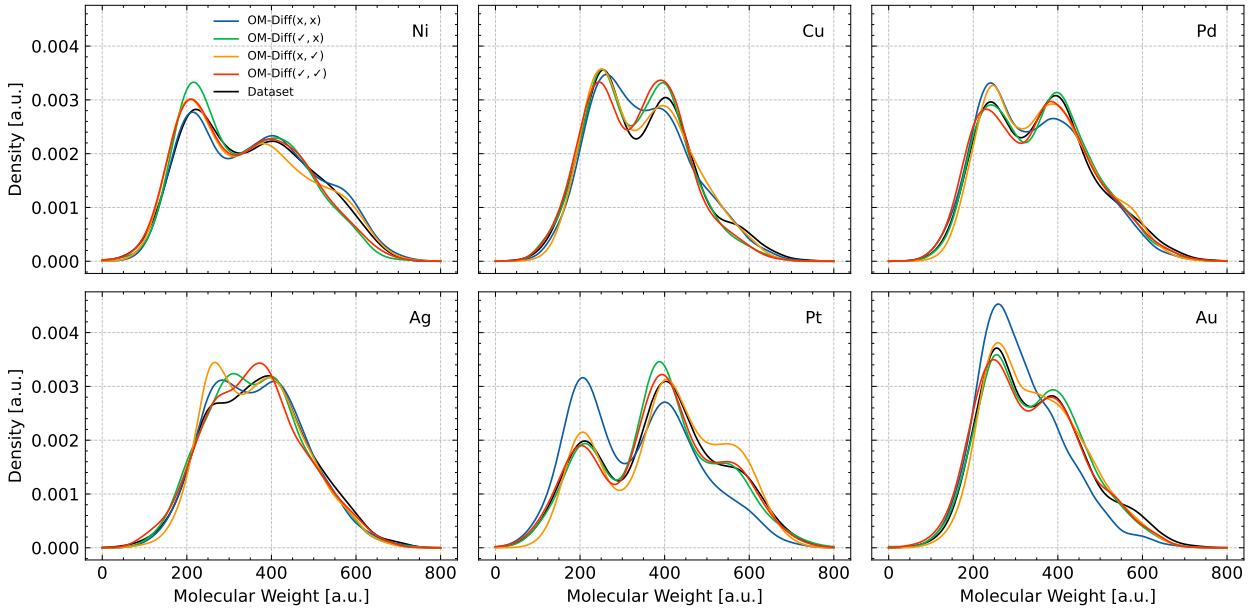


Figure S9: Distribution of molecular weight of the samples generated by the different variants of the generative diffusion model, detailed for each model. The mass of the metal center has been subtracted such that only the weight of the ligands is displayed.

Table S7: Mean Absolute Error (MAE) of the surrogate used for final screening detailed for each metal center. The presented values are given in kcal/mol. **Lower is better.** For baselines SLATM[26] and string+MLP[42], we report results are from the respective papers.

	μ_M	SLATM[26]	string+MLP[42]	Ours (MSE)	Ours (revHuber)
All	6.49	2.61	2.42	2.14 ± 0.08	2.04 ± 0.08
Ni	7.69	3.74	–	3.85 ± 0.36	3.84 ± 0.47
Cu	7.32	4.04	–	2.64 ± 0.25	2.53 ± 0.19
Pd	7.12	2.81	–	2.07 ± 0.14	1.94 ± 0.12
Ag	5.12	2.08	–	2.06 ± 0.44	1.91 ± 0.36
Pt	7.72	1.81	–	1.77 ± 0.28	1.63 ± 0.24
Au	4.28	1.60	–	1.49 ± 0.17	1.44 ± 0.19

S3 Evaluation of the surrogate models

In this section, we provide details of the evaluatuon of the different surrogate models employed in this work.

S3.1 Screening surrogate

In this section, we detail the performance of the screening surrogate for the two different loss functions employed in this work. For each variant, we performed 10-fold cross validation in order to get an error estimate for each sample in the training database.

In Fig. S10, we display the residuals. Fig. S11 shows the MAE across the property space, while Tables S8 to S10 provide the numerical details, comparison against baselines.

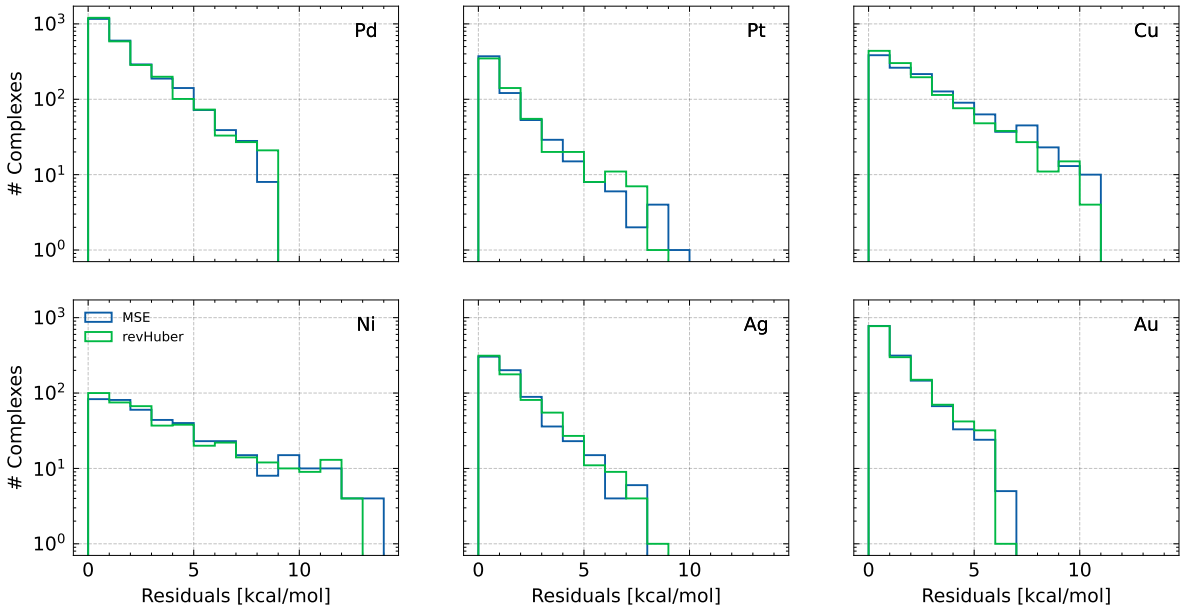


Figure S10: Residuals of the two variants of loss functions employed to train the screening regressor for each metal center. 'MSE' refers to mean-square error, while 'revHuber' stands for reverse Huber.

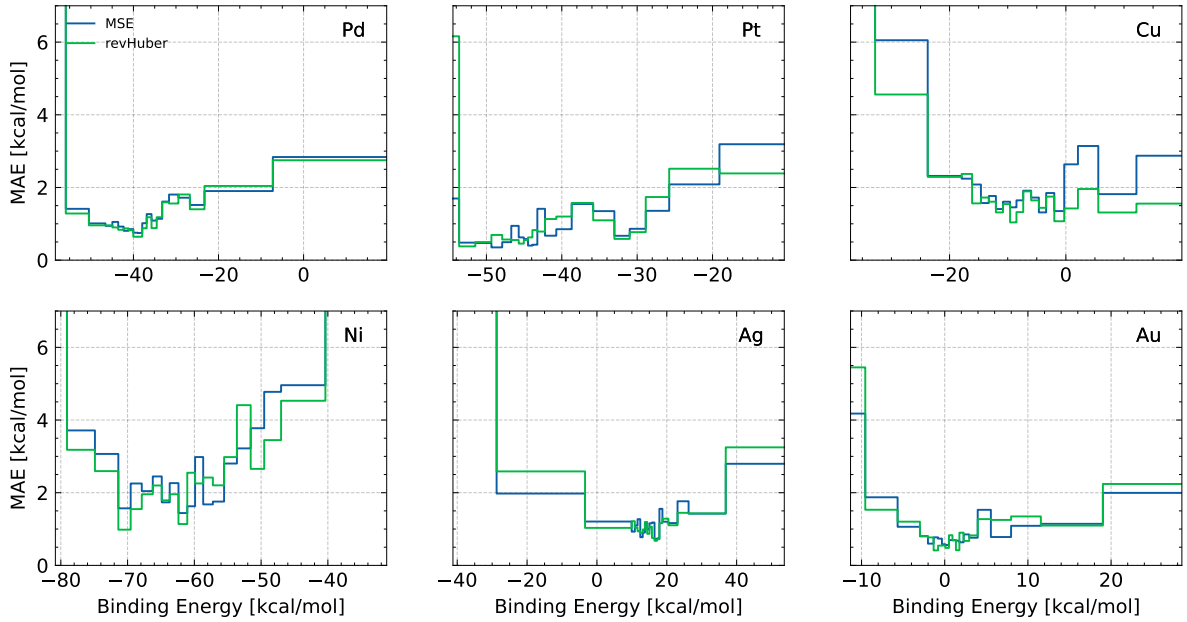


Figure S11: MAE across the property space of the two variants of loss functions employed to train the screening regressor, shown for each metal center. 'MSE' refers to mean-square error, while 'revHuber' stands for reverse Huber.

Table S8: Root Mean Square Error (RMSE) of the surrogate used for final screening detailed for each metal center. The presented values are given in kcal/mol. **Lower is better.** For baseline string+MLP[42], we report results from the paper.

	μ_M	string+MLP[42]	Ours (MSE)	Ours (revHuber)
All	8.50	3.85	3.50 ± 0.34	3.42 ± 0.29
Ni	9.50	–	5.40 ± 0.68	5.44 ± 0.96
Cu	9.27	–	3.84 ± 0.45	3.79 ± 0.30
Pd	9.16	–	3.35 ± 0.67	3.23 ± 0.64
Ag	7.15	–	3.58 ± 1.62	3.38 ± 1.54
Pt	9.26	–	2.95 ± 0.65	2.84 ± 0.49
Au	5.92	–	2.41 ± 0.43	2.33 ± 0.37

Table S9: Maximum Absolute Error (Max AE) of the surrogate used for final screening detailed for each metal center. The presented values are given in kcal/mol. **Lower is better.** For baseline string+MLP[42], we report results are from the respective papers.

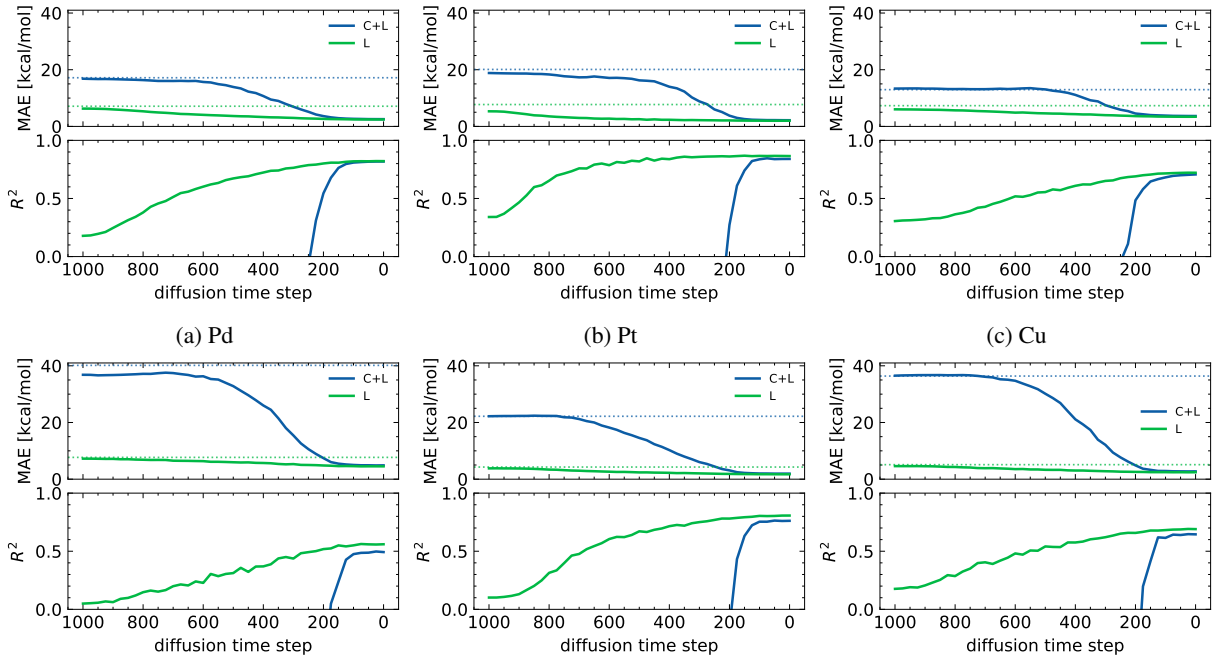
	μ_M	string+MLP[42]	Ours (MSE)	Ours (revHuber)
All	41.71 ± 12.38	26.02	32.09 ± 16.94	32.36 ± 16.37
Ni	23.79 ± 5.09	–	17.03 ± 4.13	17.33 ± 6.65
Cu	27.56 ± 2.96	–	16.90 ± 6.74	17.49 ± 6.50
Pd	35.84 ± 12.64	–	21.83 ± 15.97	21.45 ± 15.33
Ag	26.21 ± 13.90	–	18.77 ± 15.48	17.96 ± 14.88
Pt	24.25 ± 4.14	–	12.35 ± 3.60	12.89 ± 3.43
Au	22.45 ± 4.03	–	13.60 ± 6.38	12.51 ± 5.08

Table S10: Coefficient of determination (R^2) of the surrogate used for final screening detailed for each metal center. The presented values are given in kcal/mol. **Higher is better.** For baseline string+MLP[42], we report results are from the respective papers.

	string+MLP[42]	Ours (MSE)	Ours (revHuber)
All	0.974	0.978 ± 0.004	0.979 ± 0.004
Ni	–	0.652 ± 0.104	0.650 ± 0.104
Cu	–	0.824 ± 0.043	0.830 ± 0.027
Pd	–	0.865 ± 0.043	0.874 ± 0.041
Ag	–	0.735 ± 0.161	0.763 ± 0.149
Pt	–	0.891 ± 0.046	0.899 ± 0.039
Au	–	0.830 ± 0.050	0.842 ± 0.036

S3.2 Time-conditioned surrogate

In Fig. S12, we display the error of the time-conditioned surrogates detailed for each metal center.



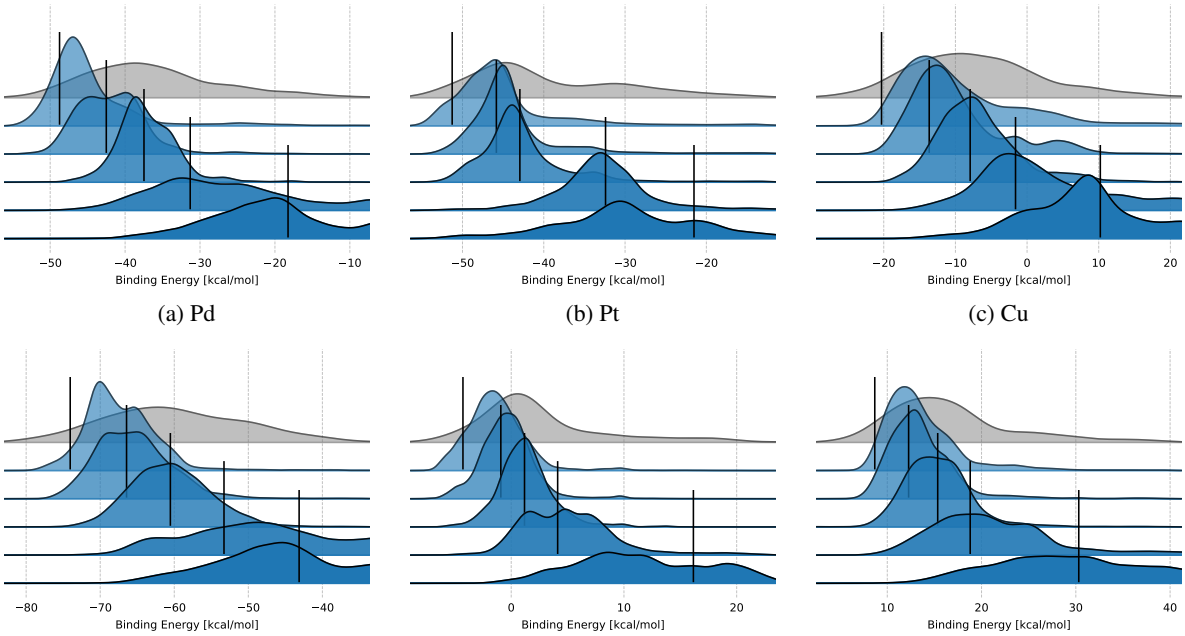
(a) Pd (b) Pt (c) Cu

(d) Ni (e) Au (f) Ag

Figure S12: Performance of the two variants of the time-conditioned regressor as a function of the diffusion time step for each metal center individually. C+L refers to the noise model that jointly corrupt center and ligands, whereas L stands for the noise where the corruption is limited to the ligands. The horizontal dotted lines represent the errors of the mean and conditional mean predictors.

S4 Conditional sampling

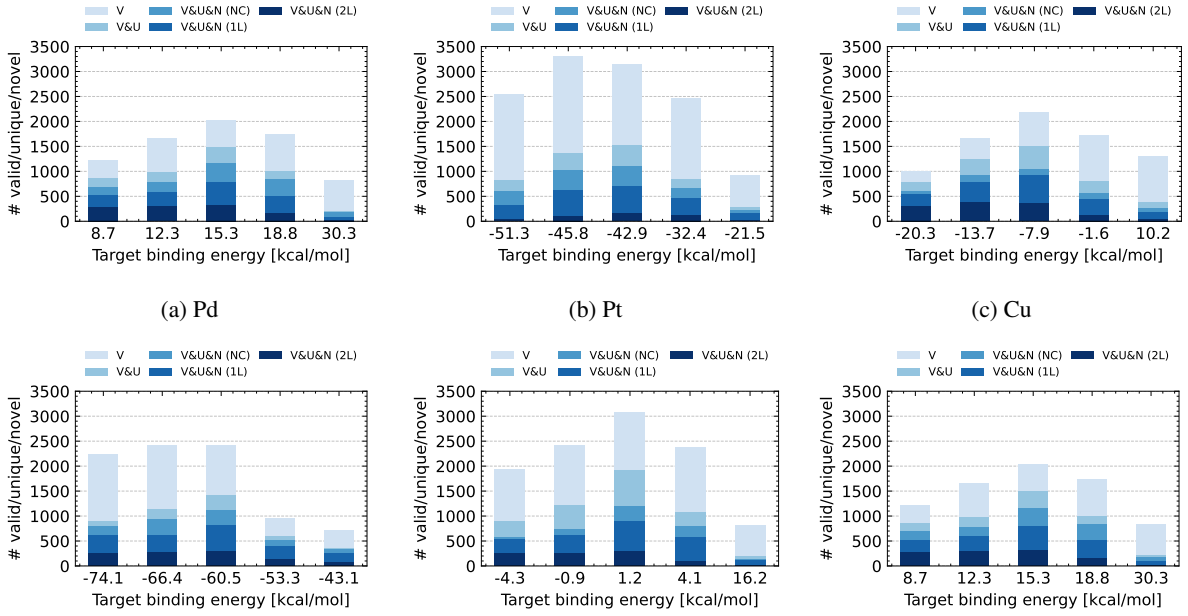
In this section, we provide the conditional distributions for metal center in Fig. S13, and the corresponding Validity/Uniqueness/Novelty breakdowns in Fig. S14.



(a) Pd (b) Pt (c) Cu

(d) Ni (e) Au (f) Ag

Figure S13: Binding energy distributions, as evaluated by surrogate, obtained through conditional sampling of OM-DIFF. The distribution in grey in the background represents the training data distribution, i.e. DFT labels. Black vertical lines represent target values.



(a) Pd (b) Pt (c) Cu

(d) Ni (e) Au (f) Ag

Figure S14: #Valid, #(Valid & Unique) and #(Valid & Unique & Novel) complexes for conditionally sampled complexes. The novelty is further divided in 3 categories: 'NC' standing for 'Novel Combination', '1L' referring to samples where 1 ligand is novel, and '2L' referring to samples where both ligands are novel.

S5 Overview of dataset

In Fig. S15, we provide an overview of the different metal-ligand combinations that are found in the dataset. All ligands are illustrated in Figs. S16 and S17.

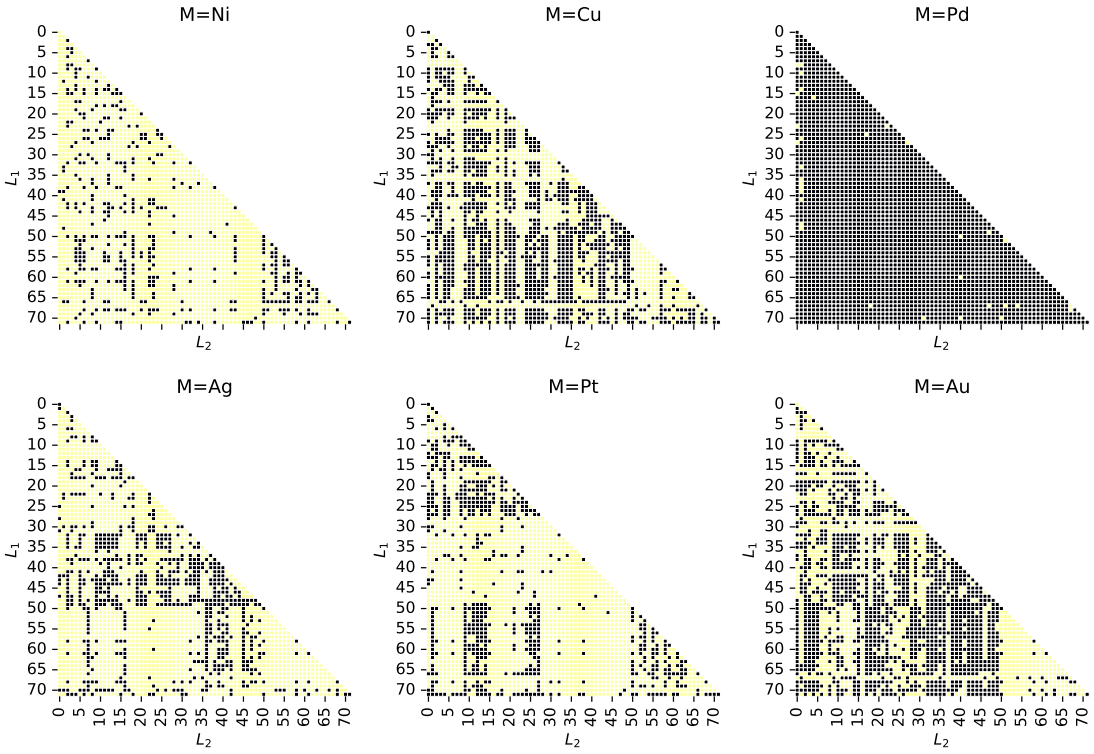


Figure S15: Combinations metal-ligand $L_1 - M - L_2$ composing the dataset. Black squares represent data points present in the dataset.

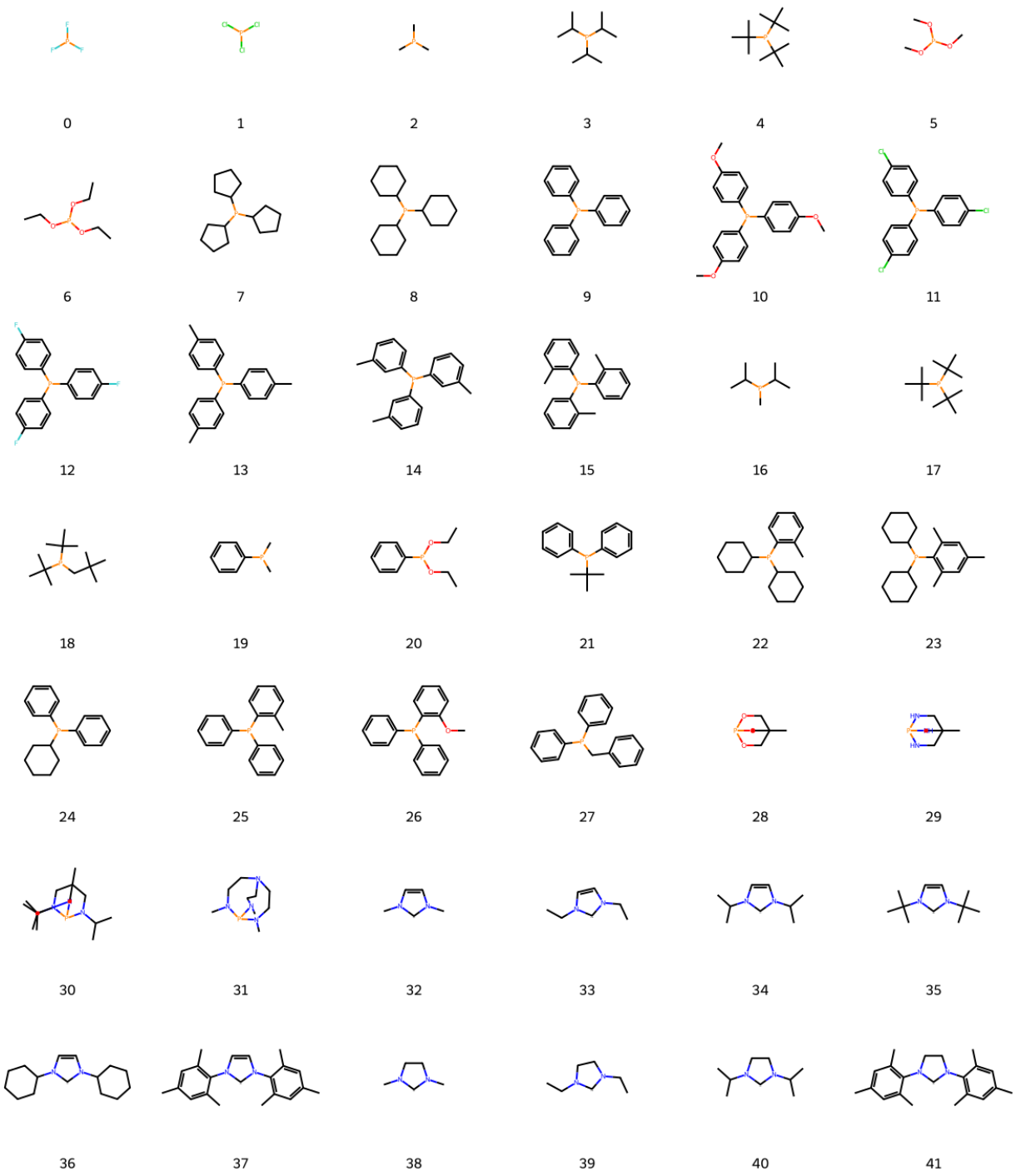


Figure S16: Ligands 0-41 used to build the dataset[26].

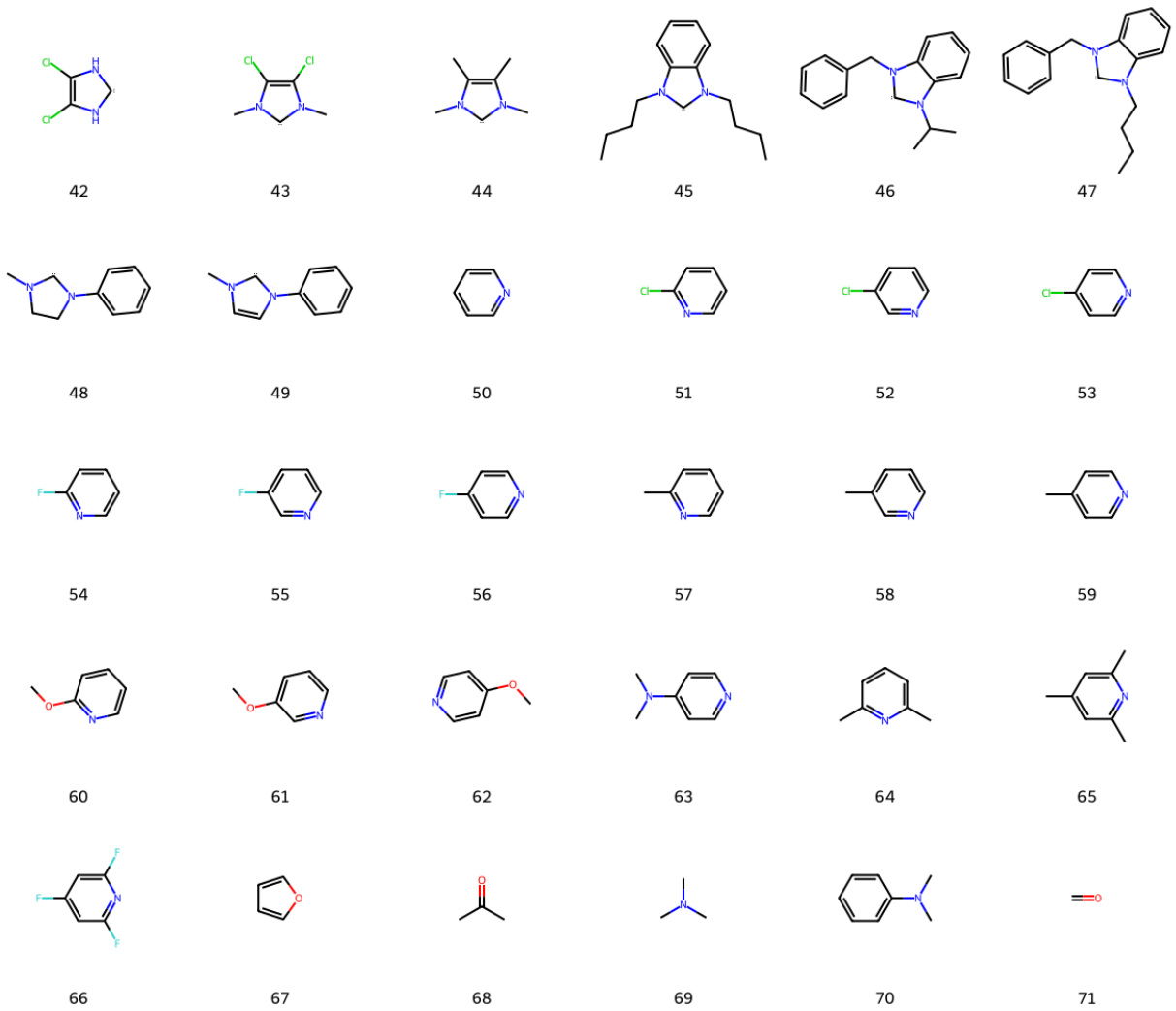


Figure S17: Ligands 42-72 used to build the dataset[26].