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

A PREPRINT

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

March 27, 2024

ABSTRACT

Organometallic complexes are ubiquitous in homogeneous catalysis and other technological applications. Optimization of such complexes for specific applications is challenging due to the large variety of possible metal-ligand combinations and ligand-ligand interactions. Here we present OM-DIFF, an inverse design framework based on a diffusion generative model for *in-silico* design of such complexes from scratch. Given the importance of the spatial structure of a catalyst, the model directly operates on all-atom (including H) representations in 3D space. To handle the symmetries inherent to that data representation, OM-DIFF combines an equivariant diffusion model and an equivariant property predictor to drive sampling at inference time. The model can conditionally generate novel ligands beyond those in the training dataset. We demonstrate the potential of the proposed approach by designing catalysts for a family of cross-coupling reactions, and validating a selection of novel proposed compounds with DFT calculations.

Keywords Generative Modelling · Denoising diffusion · Organometallic complexes · Inverse design · Cross coupling reactions

1 Introduction

In-silico catalyst design is a grand chemical challenge [1, 2], and the combination of machine learning (ML) and quantum chemistry (QC) methods is an appealing strategy to tackle it. Chemical space can be searched for useful molecules in a number of ways [3]. In a library based screening approach, inexpensive surrogate ML models can be trained with reference databases [4] and used to speed up property evaluation while preserving the accuracy of the reference method [5]. In a more recent direction [6], generative ML models can be used to learn the molecular structure distribution (often jointly with property labels) of the chemical space of interest, and in turn generate novel chemical structures that share aggregate properties with the training data (*distribution learning*) [7]. The combination of surrogate

and generative modelling opens the door to the inverse-design of materials and molecules with optimised properties, i.e. *goal-directed generation* [8]. Beyond the design of catalysts discussed herein, this approach is promising for inverse design of metal organic frameworks [9, 10], battery materials [11], photovoltaics [12], as well as drug molecules [13].

Inverse catalyst design with generative ML involves defining the relevant catalyst chemical space for the reaction of interest, as well as collecting an adequate amount of QC data for model training. While directly learning a conditional generative model is a possibility with a sufficient number of labeled samples with the desired properties, curating such dataset is computationally expensive with high accuracy QC methods. Instead, *guidance* decouples the generative process from the conditional information, by using property information only at inference time to steer the generative model towards the target properties. Larger relevant molecular structure databases [14, 15, 16] can then be used to train the generative model, and a limited amount of task-specific labeled data (e.g. energy barriers for a particular reaction) can be enough for the surrogate model to reach satisfactory accuracy.

Homogeneous catalysts are molecular in nature. Inverse molecular design has been tackled before and various generative models have been employed. They mainly differ by the data representation they operate on and the generative paradigm. On string-based representations, seminal work includes MOLGAN [17], GRAPHRNN [18], or JT-VAE [19]. Generative modeling is however not limited to approaches that involve machine learning. Other notable examples are methods that combine atoms or fragments using tailored building rules [20], either through random search or by evolving a set of candidates in a genetic algorithm (GA) [21]. Coupled with a fast and reliable fitness function, these methods have proven very effective [22, 23].

The geometry of a catalyst is important for effective catalytic activity. Generative models that operate on molecular graphs or string representations lack information about the 3D structure. A given molecular graph, or string, can potentially correspond to multiple structurally different molecules with greatly varying properties. Additionally, bonding information is not properly defined for complexes involving transition metals, requiring non-standardized descriptors [24]. Furthermore, it is also important how a ligand binding point to the metal is represented.

Recently, generative models for 3D atomistic structures [25, 26, 27] have become competitive to geometry-free models, and the diffusion paradigm [28, 29, 26] is particularly promising. An alternative approach, the variational autoencoder (VAE), is not practical for generation of atomistic point clouds as the computation of the reconstruction term involves an expensive graph matching procedure, when using a latent space invariant to permutation and orientation. Another argument favoring diffusion is the expressivity that results from mapping the prior distribution to the data distribution through a series of (simple) transitions with shared parameters, whereas standard VAEs generate samples in a single shot. Working directly in 3D also allows for leveraging advances in neural network force fields. Since the first article that demonstrated equivariant diffusion for molecules [26], multiple further developments have been done to tackle various problems such as conformer generation [30], linker design [31], structure-based design [32], or target-aware design [33]. Guidance towards target properties using an energy function [34] has also been employed, in an effective inverse-design framework operating on fragments [35].

Organometallic complexes are a challenging but high value target for molecular inverse design due to their immense popularity as molecular catalysts along with other technologically important applications such as drugs, sensor device, photonic materials, specialty polymers etc. In this work, we introduce OM-DIFF an inverse-design framework based on a guided equivariant denoising diffusion model specifically designed to generate 3D structures of organometallic complexes with optimized properties. An overview of the proposed framework is presented in Fig. 1. We summarise our main contributions as follows:

- We implement a 3D equivariant diffusion generative model, specifically designed for organometallic complexes;
- We train an equivariant property predictor and use it in combination with the diffusion model to perform regressor-guidance, and sample organometallic catalysts with targeted properties;
- We analyze several key design choices needed to attain a practical performance level, such as treating the metal center as contextual information, and varying the expressivity of the denoiser architecture;
- For the specific problem of optimizing a critical step in cross-coupling reactions, we close the loop by validating a selection of generated complexes using DFT calculations, and identify several novel complexes of potential interest.

2 Methods

2.1 Data representation

In a computer, molecular complexes are commonly represented as unordered point clouds of atoms embedded in Euclidean space. Each atom is assigned a position vector, an atom type, and potentially other features such as charges.

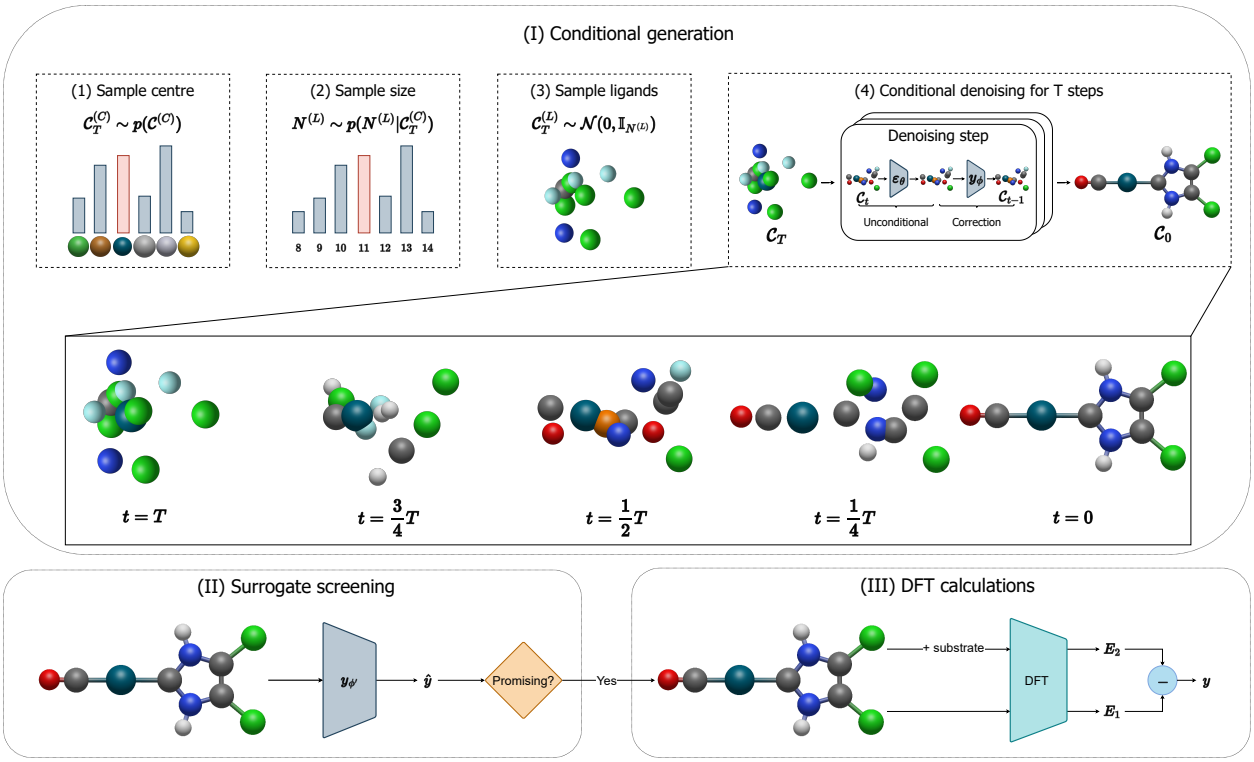


Figure 1: Inverse-design workflow. (**Top**) Overview of the conditional generation process of an organometallic complex $\mathcal{C}$: (1) A metal center, $\mathcal{C}_T^{(C)}$, is sampled; (2) based on the centre, the number of atoms contained in the coordinated ligands is sampled, (3) random atomic types and positions are assigned to all ligand atoms, and (4) the conditional denoising runs for T steps. Each denoising step involves an unconditional denoising update (steering towards valid molecules, via ε_θ) followed by a property target correction (steering towards molecules with the desired properties, via y_ϕ). The inset shows an example of a denoising trajectory for a complex with a Pd center. The position and atomic type of the center are kept fixed during the whole trajectory, and only the surrounding atoms are denoised. Their positions and types are allowed to change over the course of the generation. (**Bottom**) After a validity check, the generated complexes are screened using a surrogate model $y_{\phi'}$. The promising complexes are furher validated with DFT calculations. In the experiments of this paper, the property of interest is an energy difference, vide infra in Section 3.1.

Organometallic complexes are typically composed of a center, made of one or more transition metals, surrounded by organic ligands coordinated in specific ways. Based on that observation, we represent an organometallic compound, $\mathcal{C}$, by two distinct subsets: one with the atoms belonging to the center, denoted $\mathcal{C}^{(C)}$, and the other with the atoms belonging to the ligands, denoted $\mathcal{C}^{(L)}$. Formally, we write

$$\mathcal{C} = \{\mathcal{C}^{(C)}, \mathcal{C}^{(L)}\} = \{[x^{(C)}, h^{(C)}], [x^{(L)}, h^{(L)}]\}, \tag{1}$$

where $x^{\{(C),(L)\}} \in \mathbb{R}^{\{N_C, N_L\}\times 3}$ represents the atomic coordinates, and $h^{\{(C),(L)\}} \in \mathbb{R}^{\{N_C, N_L\}\times M}$ the atom types.

Modelling metal center and ligands separately is motivated by two factors. On one hand, the geometry of the region around the center often has to follow strict (known) rules. For instance, the square planar geometry is prevalent for transition metal complexes with d^8 configuration. On the other hand, when generating novel catalysts one wants to have full control over the center in order, for example, to enforce the central atom to be an earth-abundant transition metal. Depending on the problem under study, the positions ($x^{(C)}$), and/or the compositions ($h^{(C)}$), or possibly parts thereof, can therefore be fixed and viewed as a form of context. Scaffold-based design can also be performed by including some of the ligands in the center subset.

Next to the structural information, we often have a database of properties of interest associated with each complex. These can for instance be energies, polarizability, or dipole moment magnitude. We denote by $y \in \mathbb{R}$ the property of interest associated with a complex.

2.2 Diffusion model for organometallic catalysts

Our generative model is a tailored extension of the equivariant diffusion model for atomistic point cloud [26]. The objective of the diffusion model is to learn an unconditional distribution over organometallic complexes, $p(\mathcal{C})$, from which we can readily sample novel complexes.

Diffusion models [28, 29] are generative models that include two distinct processes: (1) a fixed *diffusion process* that iteratively corrupts data points (atomistic structures) towards a known prior distribution through additive noise, and (2) a generative *denoising process* optimised to reverse the diffusion process. The denoising process is usually learned through a neural network, that in turn can be used to sample novel atomistic structures starting from complete noise.

Diffusion process The forward diffusion gradually destroys data points, through T steps of a noisy Markov process, ending up in complete noise. Formally, this procedure defines a distribution over T latent variables $\{\mathcal{C}_t\}_{t=1}^T$, that can be seen as increasingly noisy versions of an initial atomistic structure $\mathcal{C}$,

$$q(\mathcal{C}_1, ..., \mathcal{C}_T|\mathcal{C}) = q(\mathcal{C})q(\mathcal{C}_1|\mathcal{C})...q(\mathcal{C}_T|\mathcal{C}_{T-1}). \tag{2}$$

Each transition is Gaussian, and is defined as

$$q(\mathcal{C}_t|\mathcal{C}_{t-1}) = \delta\Big([\mathcal{C}_{t-1}^{(C)}, \mathcal{C}_t^{(C)}]\Big) \cdot \mathcal{N}\Big(\mathcal{C}_t^{(L)}\Big|\sqrt{1-\beta_t}\mathcal{C}_{t-1}^{(L)}, \beta_t\mathbb{I}\Big) \tag{3}$$

where $0 < \beta_t < 1$ is some noise schedule that specifies how much information is destroyed, and $\delta(\cdot)$ is the Dirac delta measure. Intuitively, Eq. (3) means that, at each transition, the ligand features are destroyed by (1) being scaled down by $\sqrt{1-\beta_t}$, and (2) being summed with Gaussian noise of variance β_t. The center is left unchanged. For large enough T, the terminal distribution of the ligand features (i.e. positions and atom types) becomes data-independent, $q(\mathcal{C}_T^{(L)}) \approx \mathcal{N}(0, \mathbb{I})$. Due to the formulation of the diffusion process in Eqs. (2) and (3), i.e. Markovianity and Gaussian transitions, any time marginal, or said otherwise the distribution of any $\mathcal{C}_t$ given $\mathcal{C}$, can be derived analytically as

$$q(\mathcal{C}_t|\mathcal{C}) = \delta\Big([\mathcal{C}^{(C)}, \mathcal{C}_t^{(L)}]\Big) \cdot \mathcal{N}\Big(\mathcal{C}_t^{(L)}\Big|\alpha_t\mathcal{C}^{(L)}, \sigma_t^2\mathbb{I}\Big), \tag{4}$$

where $\alpha_t = \sqrt{1-\sigma_t^2}$, and σ_t is a function of the noise schedule $\{\beta_{t'}\}_{t'=1}^t$ up to time t. Through reparametrisation, any noisy version of $\mathcal{C}$, $\mathcal{C}_t$, can then be obtained without the need of going through the whole chain defined in Eq. (2),

$$\mathcal{C}_t = \Big\{\mathcal{C}^{(C)}, \alpha_t\mathcal{C}^{(L)} + \sigma_t\epsilon\Big\}, \tag{5}$$

where $\epsilon \sim \mathcal{N}(0, \mathbb{I})$. This formulation revels particularly useful for learning the generative denoising process.

Denoising process We seek to learn a denoising process that *reverses* Eq. (3), i.e. that can denoise $\mathcal{C}_t$ to $\mathcal{C}_{t-1}$. In what follows, we simplify notations by omitting the Dirac distribution for the center. When we have access to $\mathcal{C}$, the true denoising process is another normal distribution that writes

$$p(\mathcal{C}_{t-1}^{(L)}|\mathcal{C}, \mathcal{C}_t) = \mathcal{N}\Big(\mathcal{C}_{t-1}^{(L)}\Big|\mu_{\mathcal{C}_{t-1|t}^{(L)}}, \sigma^2_{\mathcal{C}_{t-1|t}^{(L)}}\mathbb{I}\Big), \tag{6}$$

where the mean and variance are given by

$$\mu_{\mathcal{C}_{t-1|t}^{(L)}} = \frac{\alpha_{t-1}}{\alpha_t}\frac{\sigma_{t-1}^2}{\sigma_t^2}\mathcal{C}_t^{(L)} + \Big(\alpha_{t-1} - \frac{\sigma_{t-1}^2}{\sigma_t^2}\frac{\alpha_t^2}{\alpha_{t-1}}\Big)\mathcal{C}^{(L)}, \tag{7}$$

$$\sigma^2_{\mathcal{C}_{t-1|t}^{(L)}} = \sigma_{t-1}^2 - \frac{\sigma_{t-1}^4}{\sigma_t^2}\frac{\alpha_t^2}{\alpha_{t-1}^2}. \tag{8}$$

While Eq. (8) only depends on the (known) time schedule, Eq. (7) involves $\mathcal{C}^{(L)}$ that is unknown at sampling time, as it is the structure we seek to generate. Using Eq. (5), we can nevertheless rewrite $\mathcal{C}^{(L)} = \frac{1}{\alpha_t}(\mathcal{C}_t^{(L)} - \sigma_t\epsilon)$, such that $\mathcal{C}^{(L)}$ can now be determined given its current noisy version $\mathcal{C}_t^{(L)}$, and the noise ϵ. The latter is not known exactly but can be approximated using a neural network ε_θ trained to map $\mathcal{C}_t$ to ϵ. We can then parametrise our generative model using ε_θ, as

$$p_\theta(\mathcal{C}_{t-1}^{(L)}|\mathcal{C}_t) = \mathcal{N}\Big(\mathcal{C}_{t-1}^{(L)}\Big|\mu_\theta(\mathcal{C}_t, t), \sigma^2_{\mathcal{C}_{t-1|t}^{(L)}}\mathbb{I}\Big), \tag{9}$$

where the variances comes from Eq. (8), and the mean is expressed as

$$\mu_\theta(\mathcal{C}_t, t) = \frac{\alpha_{t-1}}{\alpha_t}\Big(\mathcal{C}_t^{(L)} - (\sigma_t - \frac{\sigma_{t-1}^2}{\sigma_t}\frac{\alpha_t^2}{\alpha_{t-1}^2})\varepsilon_\theta(\mathcal{C}_t, t)\Big). \tag{10}$$

Unconditional sampling procedure Once trained, ε_θ can be used to generate novel samples. The unconditional sampling procedure is outlined as follows: (1) we sample the center $\mathcal{C}^{(C)} \sim p(\mathcal{C}^{(C)})$ – possibly composed of multiple atoms, by drawing it from an empirical distribution over centers[1]; (2) we then sample the number of remaining atoms that will compose the ligands given the center, $N_L \sim p(N_L|\mathcal{C}^{(C)})$; (3) we sample the initial positions and atom types of the ligand atoms, $\mathcal{C}_T^{(L)} \sim \mathcal{N}(0, \mathbb{I}_{N_L})$; (4) we finally employ the trained denoising neural network ε_θ, and execute the standard ancestral sampling procedure by iteratively applying Eq. (9).

Task and training procedure The denoising neural network ε_θ is presented with noisy atomistic structures, $\mathcal{C}_t$, obtained using Eq. (5), and is tasked with predicting the noise ϵ that was sampled to obtain the corrupted structures. The training objective follows naturally, and the parameters of the denoising neural networks are optimised to minimise the so-called simplified loss objective [36],

$$L(\theta) = \mathop{\mathbb{E}}_{\substack{\mathcal{C}\sim q(\mathcal{C}) \\ t\sim\mathcal{U}([1,\dots,T]) \\ \mathcal{C}_t\sim q(\mathcal{C}_t|\mathcal{C})}} \Big[||\epsilon_{x^{(L)}} - \hat{\epsilon}_{x^{(L)}}||^2 + ||\epsilon_{h^{(L)}} - \hat{\epsilon}_{h^{(L)}}||^2\Big], \tag{11}$$

where $[\hat{\epsilon}_{x^{(L)}}, \hat{\epsilon}_{h^{(L)}}] = \varepsilon_\theta(\mathcal{C}_t, t)$ denotes the output of the denoising network, and $[\epsilon_{x^{(L)}}, \epsilon_{h^{(L)}}] \sim \mathcal{N}(0, \mathbb{I})$ is the noise sampled to form $\mathcal{C}_t$ according to Eq. (5). We also note that, in addition to $\mathcal{C}_t$, the denoising neural network ε_θ is also provided with the time step t. While not strictly needed, providing the time step helps learn a time-dependent function more easily.

Symmetries of the learning problem Due to the geometric nature of atomistic structures, an appropriate neural network architecture should be used. The learning problem defined in Eq. (11) features different symmetries that should be accounted for. First, as atomistic structures have a set structure, i.e. their atoms feature no intrinsic order, they require ε_θ to be permutation-equivariant. This requirement intuitively means that permuting the order of the input atoms should result in a similar permutation of the output of the neural network. Second, ε_θ has to be invariant to translations of the input. In other words, the output of the network should not depend on the geometric center of the input atomistic structure. Finally, the neural network should be equivariant to rotations and reflections of the input atomistic structure. In other words, rotating / reflecting the input structure should lead to an identical rotation / reflection of the output. Formally, the two last desirata (translation invariance and rotation equivariance) write

$$\varepsilon_\theta\big([R_x x + t_x, h], t\big) = [R_x \hat{\epsilon}_{x^{(L)}}, \hat{\epsilon}_{h^{(L)}}] \tag{12}$$

where $t_x \in \mathbb{R}^{1\times 3}$ denotes any vector, and $R_x \in \mathbb{R}^{3\times 3}$ denotes any orthogonal (rotation/reflection) matrix.

Invariance of the learned distribution The likelihood of a given complex under the learned distribution $p_\theta(\mathcal{C})$ should be invariant under rigid transformations. In other words, all possible orientations of a given complex should have the same probability of being sampled. Rotation invariance is ensured by the combination of the rotation-equivariant architecture of ε_θ, and the isotropic Gaussian prior and transition distributions [30]. Translation invariance can be ensured by the translation-invariant architecture of ε_θ, and by having a prior and transition distributions over atomistic positions with fixed center of mass, $\sum_{i=1}^{N_C+N_L} x_i = 0$. Alternatively, translation invariance can be ensured by keeping the position of the metal center fixed, which is what we do.

Architecture of ε_θ We parametrise our denoising neural network, ε_θ, with a graph neural network that uses $E(3)$-equivariant message passing. Similar in nature to the PAINN architecture [37], a set of (equivariant) vectorial features is maintained and updated for each atom, along with the usual scalar features. Such architecture is provably more expressive than the original EGNN [38], as it can resolve local angular information [39] while remaining cheap to evaluate compared to higher-order architectures. During the message-passing phase, the scalar and vectorial atom features are updated, but we do not directly update positions within the message-passing phase, as it is done in the original equivariant diffusion [26]. The final scalar features are pooled to predict $\hat{\epsilon}_{h^{(L)}}$, while the final vectorial features are aggregated in a single vector $\hat{\epsilon}_{x^{(L)}}$. More details about the neural network architecture are given in Section S1.1.

[1] In the simplest case, this is simply a distribution over a single metal center – as further in this paper. The proposed framework is however not limited to such simple case. Depending on the problem under study, one could for instance sample the metal first, and then sample the coordination pattern conditioned on the metal, leading to a set $\mathcal{C}^{(C)}$ made of multiple atoms.

2.3 Regressor guidance

So far, we have introduced the necessary tools for learning a generative model from which novel atomistic structures can be sampled unconditionally. When performing inverse-design, we are interested in being able to sample novel atomistic structures with optimised properties, i.e. sample from $q(\mathcal{C}|y)$. In this section, we present a procedure, named regressor-guidance, for performing conditional sampling using the pretrained unconditional generative model presented in Section 2.2 combined with an auxiliary property predictor.

Conditional model Inspired by classifier-guidance [40], regressor-guidance builds on the observation that conditioning on a property of interest y can be done by sampling from a conditional denoising process

$$p_{\theta,\phi}(\mathcal{C}_{t-1}|\mathcal{C}_t, y) \propto p_\theta(\mathcal{C}_{t-1}|\mathcal{C}_t)p_\phi(y|\mathcal{C}_{t-1}), \tag{13}$$

where $p_\theta(\mathcal{C}_{t-1}|\mathcal{C}_t)$ is the unconditional denoising process from Eq. (9), and $p_\phi(y|\mathcal{C}_{t-1})$ is a conditional distribution over properties induced by a property predictor y_ϕ. Here, we define $p_\phi(y|\mathcal{C}_{t-1})$ in terms of an energy function,

$$f_\phi(y, \mathcal{C}_t) = ||y - y_\phi(\mathcal{C}_t, t)||^2, \tag{14}$$

such that $p_\phi(y|\mathcal{C}_{t-1}) \propto \exp\big(-f_\phi(y, \mathcal{C}_t)\big)$. This formula is also similar to that of loss-guided diffusion (LGD) [41], with a loss function that includes a learnt component, i.e. y_ϕ. The conditional denoising process from Eq. (13) writes as a *corrected* version of the unconditional process defined in Eq. (9),

$$p_{\theta,\phi}(\mathcal{C}_{t-1}^{(L)}|\mathcal{C}_t, y) = \mathcal{N}\Big(\mathcal{C}_{t-1}^{(L)}\Big|\mu_{\theta,\phi}(\mathcal{C}_t, y, t), \sigma^2_{\mathcal{C}^{(L)}_{t-1|t}}\mathbb{I}\Big) \tag{15}$$

where the *corrected* mean is obtained as

$$\mu_{\theta,\phi}(\mathcal{C}_t, y, t) = \mu_\theta(\mathcal{C}_t, t) - \sigma^2_{\mathcal{C}^{(L)}_{t-1|t}} \nabla_{\mu_\theta(\mathcal{C}_t,t)} f_\phi\Big(y, \mu_\theta(\mathcal{C}_t, t)\Big). \tag{16}$$

The correction is obtained by evaluating the gradient of f_ϕ with respect to the mean predicted by the unconditional model. In practice, sampling from the conditional distribution amounts to first evaluating the mean of the unconditional distribution, then evaluating the energy function in Eq. (14) using the estimated mean, and finally computing the correction expressed as the gradient of Eq. (14) with respect to $\mathcal{C}_t$.

Task and training procedure As per Eq. (14), the property predictor y_ϕ is tasked with predicting the property of interest y of structure $\mathcal{C}$, given a noisy version of it $\mathcal{C}_t$. This implies that y_ϕ should be trained on noisy structures obtained with the same diffusion process as the generative model. A natural training objective could be the mean squared error (MSE),

$$L(\phi) = \mathbb{E}_{\substack{(\mathcal{C},y)\sim q(\mathcal{C}) \\ \mathcal{C}_t \sim q(\mathcal{C}_t|\mathcal{C})}}\Big[||y - y_\phi(\mathcal{C}_t)||^2\Big]; \tag{17}$$

however, in practice the prediction task becomes very difficult for high noise levels, i.e. when $t \to T$, when structures are close to pure noise. In the following, we thus resort to the Huber loss rather than the MSE to limit the influence of the most noisy structures, while we hypothesize that a more sophisticated weighting technique for the loss could result in even better performance.

Property predictor parametrisation As y is a scalar, the learning task is inherently invariant to rigid transformations of the input atomistic structure. The property predicted by y_ϕ should not depend on the orientation nor the absolute position of $\mathcal{C}_t$. With that in mind, we parametrise our property predictor y_ϕ in Eq. (14) with a neural network, that features an encoder similar to that of the denoising neural network ε_θ, i.e. that maintains and updates a set of scalar and vectorial states for each atom. The encoder is followed by a readout layer that aggregates the final scalar atom states into a sole complex-level state using an attention mechanism. That aggregated state is in turn passed to a fully-connected neural network outputting the predicted value of y.

2.4 Screening surrogate

Once samples have been generated, it is not feasible to evaluate all of them using DFT calculations. We therefore employ a surrogate model for screening the generated samples, before running further calculations with DFT on the most promising candidates. Such surrogate could technically be obtained by leveraging the property predictor y_ϕ presented in Section 2.3: $y_\phi(\mathcal{C}_t, t = 0)$. We however found that training a separate surrogate that has only seen *clean*

samples was slightly more accurate. The screening surrogate shares the same architecture as the time-conditioned regressor, but without time input.

We are interested in having a surrogate that behaves well over the whole property space, and avoids very large errors, as they could lead to missed candidates, or unfruitful expensive DFT calculations. We therefore compare the training of the same surrogate with two different loss functions, the usual MSE loss and the reverse Huber loss (revHuber). The latter switches between L1 and L2 losses: For samples where the error is below a given threshold, the L1 loss is applied, while the L2 loss is applied to penalise larger errors.

3 Experiments and Results

Organometallic complexes are a very versatile group of materials with application areas in energy, medicine, functional materials, sensors, optical devices etc. They are designed for therapeutics as metallodrugs [42] and can be used to form specialty polymers [43], to make organic light emitting materials [44], as photovoltaic materials [45], are used in batteries [46] or as sensors [47]. Organometallic catalysts are widely used in both homogeneous (solution phase) and heterogeneous (solid phase) catalytic processes. They are crucial for many industrial chemical reactions, such as hydrogenation [48], hydroformylation [49], olefin metathesis [50], or cross-coupling reactions (e.g., Suzuki [51], Heck [52], and Stille [53] reactions).

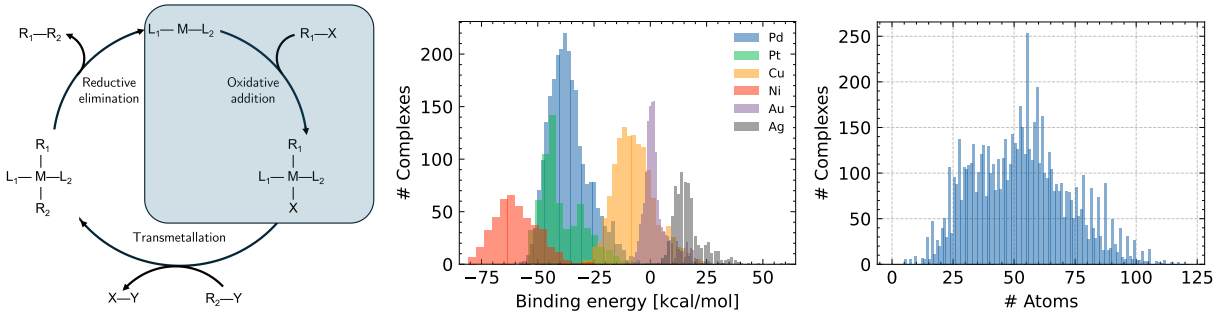
Cross-couplings reactions are fundamental in synthesizing pharmaceuticals, agro-chemicals, and organic materials. In organic chemistry, they are crucial in the synthesis of complex molecules as they enable the formation of carbon-carbon (C-C) and carbon-heteroatom (C-N, C-O, C-S, etc.) bonds efficiently. They are used for synthesizing active pharmaceutical ingredients (APIs) with diverse biological activities, organic semiconductors and conductive polymers, for creating molecular diversity by combining a broad range of substrates which is beneficial for discovering new materials and drugs. Typically cross-coupling reactions are environmentally friendly, utilize catalytic systems that minimize waste and avoid the use of toxic reagents. Specifically, the Suzuki cross-coupling reaction enables the formation of carbon-carbon bonds by coupling olefins with organoboron compounds under mild conditions. This method is specially popular due environmental friendliness, using less toxic and more stable reagents, and producing water as a byproduct, thereby aligning with principles of green chemistry. The Suzuki cross-coupling reaction is crucial for creating medicines, agricultural chemicals, electronic materials and its key importance led to the 2010 Nobel prize in chemistry [54, 55].

The Suzuki reaction is part of a larger family of cross-coupling reactions (Suzuki, Kumada, Negishi, Stille and Hiyama) [56, 57] – whose cycle is generically represented in Fig. 2a, where the leaving group is usually a halide (represented by an X in Fig. 2a). The five reactions differ by the nature of the cross-coupling partner (represented by Y in Fig. 2a). The rate limiting step of these reactions can be described with volcano plots. Although not common in homogenous catalysis, volcano plots have proved to be very useful in heterogenous catalysis. Volcano plots for the cross-coupling reactions under study have been shown to be similar [57], differing only by the width of the plateau region. The Hiyama reaction has widest plateau region spanning $[-82.2, 2.0]$ kcal/mol, while the region of interest for the Suzuki reaction is only between -32.1 and -23.0 kcal/mol. The other reactions have a plateau region lying in between that of the two aforementioned reactions. As a consequence, catalysts that bind too strongly or too weakly for the Suzuki reaction might still be efficient for other cross-coupling variants. Using volcano plots, the reaction energy of the oxidative addition process, highlighted in blue in Fig. 2a, can be used as a descriptor to estimate catalytic activity [57, 58]. This approximation allows for quicker screening of candidates than if the whole cycle needed to be computed. Relevant complexes should sit on the volcano plateau, or near the top.

As a test bench for OM-DIFF, we choose the inverse-design of organometallic catalysts for cross-coupling reactions. We choose that family of reactions for two reasons: (1) its practical relevance as argued above, (2) compelling literature showing that relevant catalysts can be searched through optimization of a binding energy that acts as a proxy for the actual catalytic activity.

3.1 Dataset

We perform our experiments using the DFT-level subset of the C–C cross-coupling database [58]. In the database, each catalyst is an inorganic complex made of a metal center that binds to two ligands, as $L_1 - M - L_2$ with $M \in \{Ni, Pd, Pt, Cu, Ag, Au\}$. The ligands L_1 and L_2 can either be identical or different. The initial database at MMFF-level [59] was constructed by combining the 6 metal centers with 91 distinct ligands. This makes for a total of 25116 possible combinations. However, optimized geometries and the corresponding oxidative addition binding energy were computed with DFT for around 7000 complexes. Those complexes cover the 6 different metal centers, and 72 unique ligands (all depicted in Figs. S16 and S17). Pd is the only metal present with all possible ligands, while the other metals are only combined with a limited number thereof. An overview of the metal-ligand combinations



(a) (b) (c)

Figure 2: **(a)** Reaction under study in this work—the model generates $L_1 - M - L_2$ and optimises the reaction energy of the oxidative addition. **(b)** Distribution of binding energy in the considered dataset [58]. **(c)** Distribution of number of atoms in the considered dataset [58]. On average complexes are composed of 53 atoms.

present in the data is provided in Fig. S15. Although DFT relaxed geometries and binding energies are available for around 7000 complexes, the diversity of the data is limited because only 72 ligands, or building blocks, were used. The binding energies are for the oxidative addition of the substrate with the transition metal, depicted in blue in Fig. 2a. The corresponding distribution across the dataset is illustrated in Fig. 2b.

DFT computational details We confirmed with DFT calculations a few promising catalyst candidates generated by OM-DIFF. We performed the calculations with the quantum chemistry software ORCA version 5.0.4 [60, 61] using a protocol similar to that of the initial study [58] that generated the training data. Specifically, we used the B3LYP functional [62, 63, 64] with the Pople 3-21G basis set [65, 66, 67, 68] for the geometry relaxation of Cu and Pd complexes and the Ahlrichs def2-SVP double-ζ basis set [69] (ORCA keyword `def2-SVP`) for Pt complexes. The parameters for the 3-21G basis set was downloaded from the basis set exchange [70]. We used the original D3 dispersion correction [71] (ORCA keyword `D3ZERO`), as per the original protocol [58]. The RIJCOSX approximation [72, 73] was used to speed up Coulomb and Exchange integrals, with the automatic generation of an auxiliary basis set [74] (ORCA keyword `autoaux`) for calculations that used the 3-21G basis set, and the def2/J auxiliary basis set [75] (ORCA keyword `def2/J`) for the def2 family of basis sets. For energy evaluations, we performed single point calculations using the Ahlrichs def2 triple-ζ basis set [69] (ORCA keyword `def2-TZVP`).

Previous work that used the data In the initial study [58], the remaining MMFF-level [59] configurations were screened using a surrogate model trained to map MMFF-level geometries to DFT-level binding energies. While promising candidates were identified, none of them was investigated further with DFT. In a generative context, a VAE operating on string representations [24] has also been applied to the dataset. The model displayed controllability, and could generate novel and promising candidates. Very recently, a GA [76] was successfully used to generate promising catalysts for the Suzuki reaction.

3.2 Unconditional generation

We first test our model for effective unconditional generation of organometallic complexes. Effective unconditional generation constitutes a prerequisite for effective conditional generation, and is also a valid inverse-design procedure when combined with screening. We evaluate and compare the ability of different ablated versions of OM-DIFF to learn the unconditional data distribution. We specifically study two aspects of the generative diffusion model: (1) the modelling of the central region as context around which the model is tasked to build the ligands, and (2) the expressiveness of the neural network architecture.

Setting After training each model variant, we generate 10000 samples, where the number of atoms is drawn from the empirical distribution displayed in Fig. 2c. Firstly, we evaluate the properties of the generated samples in terms of structural metrics: validity, uniqueness and novelty. Secondly, we compare the samples with the empirical data distribution in terms of the geometry around the metal center M, the binding energy as estimated by a surrogate model, and the composition. For the geometric metrics, we also include results from the force-field-level subset [58] to get an idea of the advantage yielded by a generation in 3D. In Fig. 6b, we additionally compare the distribution of the strain energy at XTB-level [77] of the generated compounds, i.e. energy difference between generated structure and XTB-optimised structure. Additional details about the evaluation procedure and the different baselines are provided in

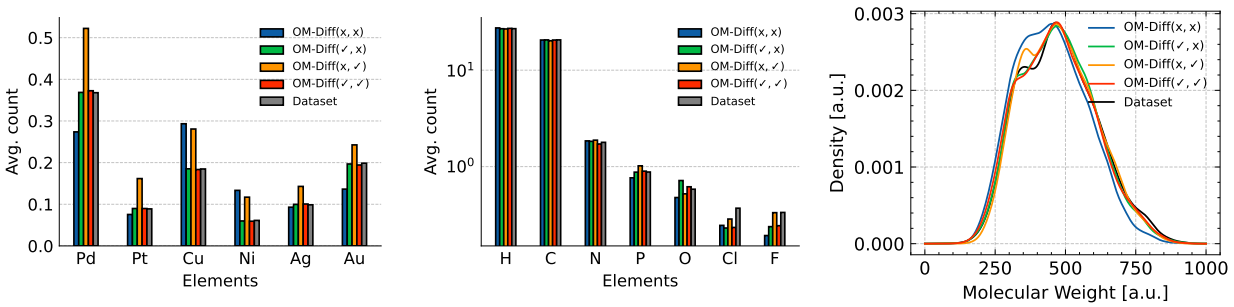
Section S2, where we additionally report the error of the diffusion variants on a held-out set as a function of the noise level Fig. S1.

Table 1: Results of the unconditional sampling experiment. Structural metrics are reported in % of the number of generated samples. Distances to the empirical data distribution are Wasserstein (W) distances for continuous features and total variation (TV) for discrete features. Details are given in S2.

	Fixed $\mathcal{C}^{(C)}$	Improved ε_θ	Structural metrics [%] (↑)			Distance to empirical distribution (↓)					
						Center-proximal geom.		Binding energy [kcal/mol]	Generated atom types [TV]		
			Valid	Valid and unique	Valid, unique and novel	Length [Å] $W^{L_{1,2}-M}$	Angle [rad] $W^{L_1-M-L_2}$	$W^{\Delta E}$	Center TV_M	Non-center $TV_{non\text{-}M}$	Center-proximal TV_{prox}
OM-Diff	x	x	8.9	7.9	4.8	0.256	0.0113	0.0044	0.347	1.036	0.079
	✓	x	19.6	16.4	7.5	0.165	0.0084	0.0044	**0.015***	**0.482**	0.029
	x	✓	18.9	17.6	11.1	0.203	0.0089	0.0040	0.388	1.005	0.041
	✓	✓	**28.2**	**23.9**	**11.8**	**0.147**	**0.0079**	**0.0036**	**0.015***	0.593	**0.019**
Merck Molecular Force Field			–	–	–	0.814	0.0257	–	–	–	–

Effect of center Modelling the center as context is beneficial for two things: (1) the geometry around the metal center, and (2) the composition. Regarding geometry, the model reproduces the training distribution of the geometry around M better, as seen from $W^{L_{1,2}-M}$ and $W^{L_1-M-L_2}$ in Table 1. Graphical depictions of the corresponding distributions can be found in Figs. S4 and S5. Regarding composition, the generated distributions over atom types are closer to the training distributions when the center is modelled as context, as hinted by TV_M, $TV_{non\text{-}M}$, and TV_{prox} (all defined in Eq. (S10)) in Table 1. TV_M is virtually zero, since the metal-center is directly sampled from the empirical distribution obtained from the training data. The detailed metal center distribution for each variant of the model can be found in Table S5. The distribution of center and non-center atomic elements are shown in Figs. 3a and 3b, whereas the distribution of center-proximal atomic elements can be found in Fig. S8 as well as the associated total variations in Table S6. Figs. 3c and S9 also indicate that models where the center is part of the context tend to generate complexes whose molecular weight (proxy for joint distribution of atom types) tend to be closer to the training data distribution.

We hypothesize that the observed improvement is due to a simplified classification task. When inspecting Fig. S1, we see lower losses on ε_h and h, especially for larger noise levels. In Table S2, we can also see that chemical validity is improved, i.e. fixing the center can make up for a less expressive backbone in terms of validity.

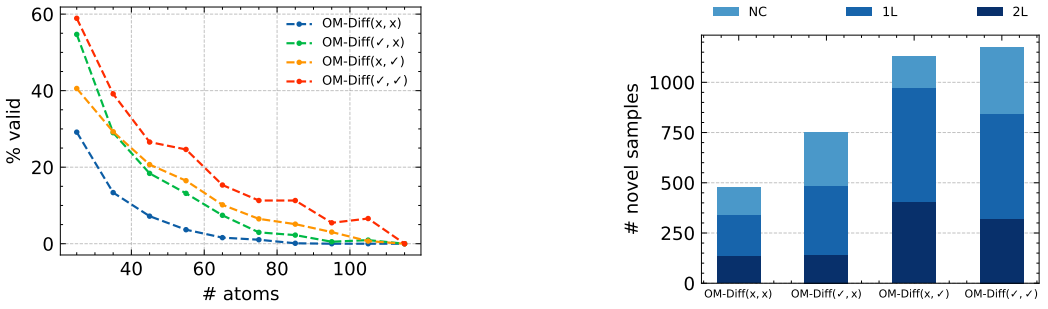


(a) (b) (c)

Figure 3: Distribution of **(a)** metal-centers, **(b)** non-metal elements, and **(c)** molecular weight in unconditional sampling.

Effect of representation We find that more expressive geometric neural networks, generally bring an increased validity, noticeably so as generated complexes get larger as highlighted in Fig. 4a. We also provide Table S2, where we can clearly see that the added expressivity yields more configurations that pass the pairwise distances check, i.e. more expressive architecture leads to less atom clashes and disconnected fragments. In Fig. S1, we can also observe that the more geometrically expressive variants yield lower losses on the ε_x and ultimately on x, i.e. the atomic coordinates. The difference is especially stronger for lower noise levels, when getting closer to the data manifold. Our findings are in line with previous work, e.g. [78], that also showed improved generative capabilities resulting from geometrically more expressive architectures.

Validity As described in details in Section S2.1, we deem a complex valid, if it features exactly one metal center, has no isolated nor clashing atoms, and passes a validity check based on the RDKIT software [79]. OM-DIFF(x, x), the



(a) (b)

Figure 4: **(a)** Validity as a function of the complex size. **(b)** Sources of novelty for the different variants of the model, where 'NC' stands for 'Novel Combination', '1L' refers to samples where 1 ligand is new, and '2L' refers to samples where both ligands are novel. For each variant, 10000 complexes were sampled.

ablated version of OM-DIFF similar to the original EDM model [26], yields only around 900 valid complexes among the 10000 generated. The two proposed modifications yield a substantial improvement, allowing a better modelling of the chemical space under study. With nearly 3000 valid structures out of 10000, this corresponds to over threefold improved efficiency of valid sample generation. In Fig. 4a, we display validity as a function of the size of the generated complexes. As expected, larger complexes tend to be more difficult to generate than smaller ones. As a molecule is deemed invalid if e.g. the valence of any of its atoms is not respected, validity at complex-level decreases quickly with complex size, even if the *atomwise* validity is kept constant. In Fig. 4a, we can clearly see that the full OM-DIFF is the only variant of the model able to generate larger valid complexes.

Uniqueness and Novelty While generating complexes that are chemically valid is an important first step, these complexes also need to be different from each other and novel. A generative model that only regenerates data it has been trained on is not extremely useful for inverse design. As seen in Table 1, all variants are able to produce varied and novel complexes. In Fig. S2, we show how Validity, Uniqueness and Novelty evolve with the number of sampled complexes. While Validity remains rather constant as the number of generated compounds is increased, Uniqueness tends to decrease, indicating that the model tends to generate given compounds multiple times. It needs to be emphasized that these metrics are obtained by converting the actual geometry into `SMILES` strings and limiting the comparison based on `SMILES` description. Two identical `SMILES` strings can actually have been obtained from (slightly) different geometries. Novelty is also observed to remain constant as the number of generated complexes is increased. Interestingly, while OM-DIFF clearly yields more valid (and unique) complexes, it is on par in terms of novelty with the variant where the metal-center is not modelled as context. Finally, when looking closer at non-unique complexes, we observe that it is mostly non-novel complexes (i.e. present in the training data) that are generated multiple times. However, the model is also able to generate novel compounds multiple times. In Fig. S3, we show the number of novel ligands generated as a function of the number of generated complexes. We can see that it steadily increases with the number of complexes generated.

Sources of novelty Due to the way the dataset was constructed, i.e. as combination of 72 ligands as detailed in Section 3.1, novelty can take different forms. Novel compounds fall into 3 different categories: (1) novel combinations of 2 existing ligands, (2) combinations of an existing ligand with a novel ligand, and (3) combination of 2 novel ligands. In Fig. 4b, we display the distribution of novelty for the different variants of the model, in percentage of the novel samples generated. We observe that variants where the center is part of the context tend to get a larger percentage of novelty coming from novel combinations, highlighting that they have learnt the data distribution (slightly) better than their counterparts that do not fix the center.

Generated complexes Given the rather limited structural diversity of the training data: 6 metal centers and 72 ligands, the model only gets to see a very restricted part of the chemical space. We have shown in the previous section that while the model can be prone to regenerate the building blocks seen during training, all model variants were able to generate novel ligands, as highlighted in Fig. 4b. We provide an excerpt of 18 novel ligands in Fig. 5 generated with the full OM-DIFF. Compared to the ones used to build the dataset, displayed in Figs. S16 and S17, we observe that the generative model tends to produce ligands in the neighbourhood of the training data, hinting that the model has learnt the underlying distribution. As a concrete example, `U9` is similar to `54` with an extra F atom.

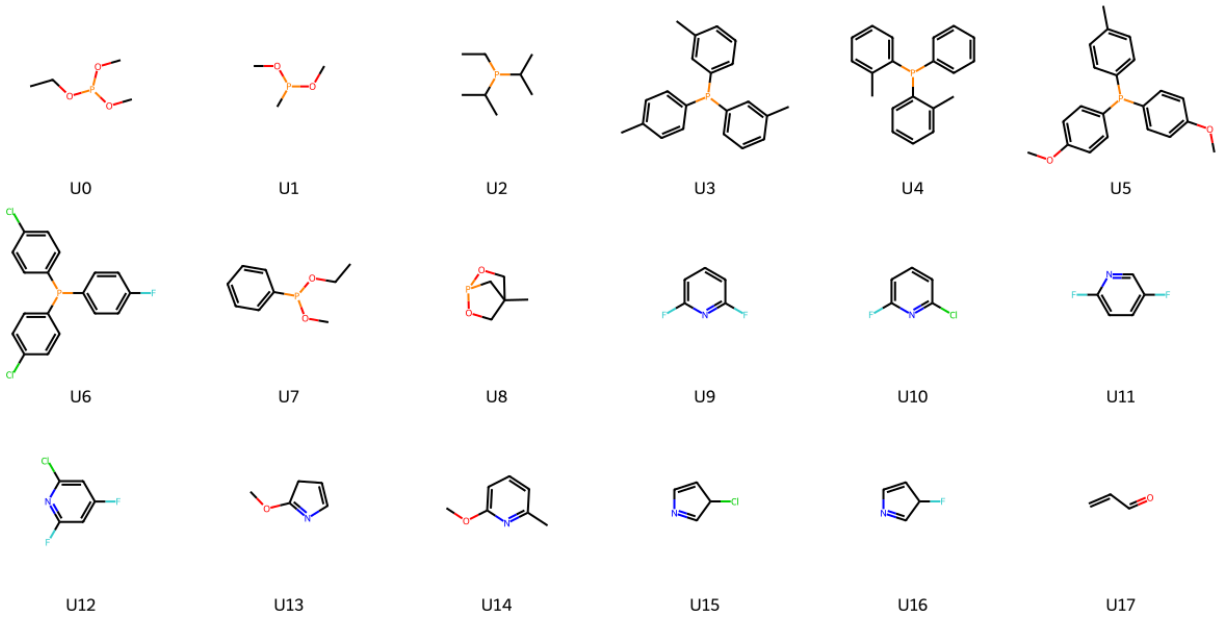
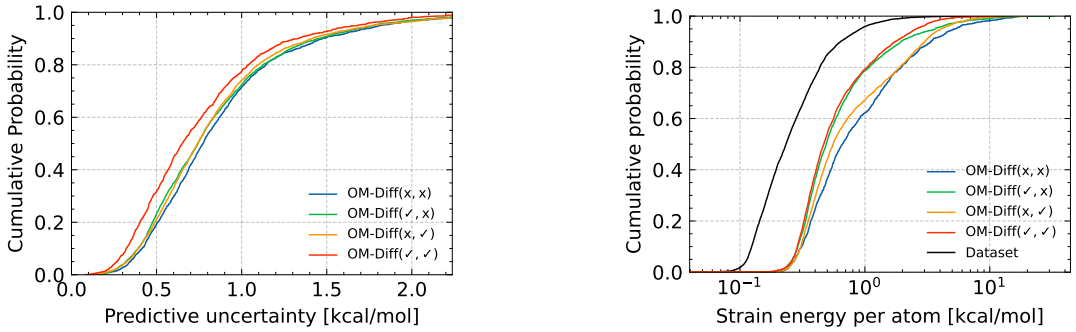


Figure 5: Novel ligands generated by OM-DIFF through unconditionalsampling. The ligands have been randomly picked, and ordered by visual similarity.

Quality of generated complexes In Table 1, we reported $W^{\Delta E}$ that quantified the discrepancy between the binding energy distribution of the generated complexes and the ground-truth one. We can additionally leverage an ensemble of screening surrogates to estimate uncertainty of the generated complexes. This is what we display in Figs. 6a and S7. The predictive uncertainty is taken as the standard deviation across 10 surrogates. While uncertainty has not been calibrated to match the actual errors, relative comparisons can still performed under the reasonable assumption that the uncertainty estimate is capable of ranking. As the different surrogate models have been trained on clean data only, we can expect them to disagree as data starts looking less and less realistic. Similar to the validity results, the full OM-DIFF model variant generates samples about which the screening surrogates disagree the least. We additionally computed the strain energy for the samples generated by the different variants of OM-DIFF. Their cumulative distributions are displayed in Fig. 6b. The strain energy per atom is defined as the energy difference between the structure generated by the generative model, and its energy after relaxation using XTB normalised by the number of atoms. For a fair comparison, we only included complexes with up to 75 atoms, as the simplest version of OM-DIFF could not generate any larger sample that was valid. Interestingly, we find variants of OM-DIFF where the center is modelled as context to generally lead lower strain energy. This can probably be explained by a better modelling of the region around the metal center, as also displayed in Table 1, via lower $W^{\mathrm{L}_{1,2}-\mathrm{M}}$ and $W^{\mathrm{L}_1-\mathrm{M}-\mathrm{L}_2}$ values.



(a) (b)

Figure 6: Cumulative distribution of (a) predictive uncertainty, as estimated by an ensemble of 10 surrogates, evaluated on the samples generated by the different variants of the generative diffusion model; (b) strain energy per atom as estimated by XTB [77].

3.3 Performance of the surrogate models

As illustrated in Fig. 1, our framework heavily relies on surrogates trained to approximate DFT-level binding energies. Such models are involved at two different steps: (1) to steer the conditional generation, and (2) to screen and filter final samples prior to DFT computations. It is therefore crucial to evaluate the accuracy of the different surrogates across the chemical space of interest. The range of the binding energy across the dataset, is around 120 kcal/mol. Following previous work, we consider a model useful if its error is within 5% of that range, i.e. around 5 kcal/mol. A dummy model that outputs the conditional mean of the dataset gets an average root-mean-squared error of around 8.3 kcal/mol. To evaluate errors, we perform a stratified 10-fold cross-validation, where folds were designed to keep the proportion of metal centers approximately equal across folds, and to cover the property range uniformly. We compare against two baselines: one based on SLATM [58], and another that trains a neural network surrogate in the latent space of a VAE [24].

Loss function To reduce large errors on outlying data points, we experimented with the reverse Huber loss (coined 'revHuber') for the screening surrogate. We observed a slight improvement in performance for model trained with the 'revHuber' loss, as illustrated in Figs. 7a and S10. We also summarize the error diagnostic of our screening surrogate in Tables 2 and S7 to S10. While our model is accurate on average, slighty more than the compared baselines, and within the 5 kcal/mol, we still observe large errors on some specific outliers. In Fig. S11, we additionally show the surrogate error across the property space for both loss functions. The different curves display a U-shape, highlighting that the surrogate predictive accuracy decreases as we get closer to the tails of the training distribution. For instance, in the case of Cu the MAE of the surrogate is above 5 kcal/mol in the plateau region of the Suzuki reaction. This has implications when looking for compounds at the boundaries of the training distribution – predictions in that area should be used cautiously.

Table 2: Validation of the surrogate used for final screening. We report mean and standard deviation across 10-fold cross-validation of the mean absolute error (MAE), root mean squared error (RMSE), 95th quantile absolute error (Q95 AR), maximum absolute error (Max AE), and coefficient of determination (R^2). Results for SLATM [58] and string+MLP [24] are obtained from the respective papers.

	MAE [kcal/mol] (↓)	RMSE [kcal/mol] (↓)	Q95 AE [kcal/mol] (↓)	Max AE [kcal/mol] (↓)	R^2 [-] (↑)
μ_M	6.49	8.50	–	41.71 ± 12.38	–
SLATM [58]	2.61	–	–	–	–
string+MLP [24]	2.42	3.85	–	26.02	0.974
Ours (MSE)	2.14 ± 0.08	3.50 ± 0.33	6.98 ± 0.31	32.09 ± 16.95	0.978 ± 0.004
Ours (revHuber)	2.04 ± 0.08	3.42 ± 0.29	6.92 ± 0.45	32.36 ± 16.37	0.979 ± 0.004

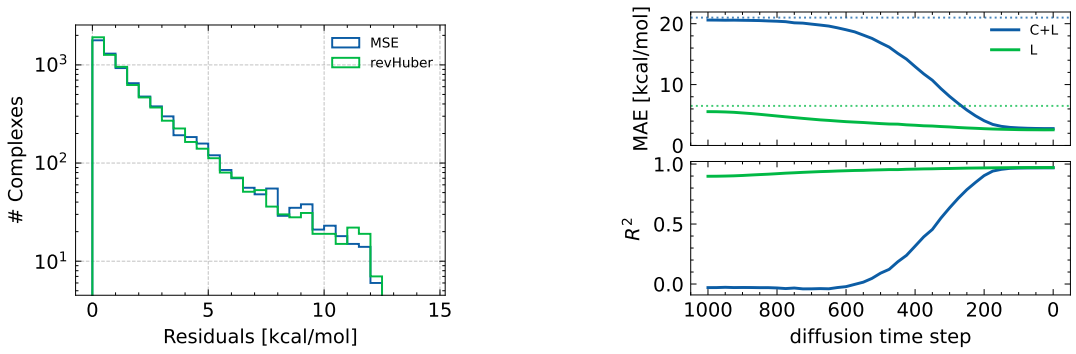
Time-conditioned surrogate In Figs. 7b and S12, we display the error of the time-conditioned regressor for the two different variants of corruption, i.e. whether the center is part of the context (variant `L`) or not (variant `C+L`). In terms of error, models tend to behave similarly to dummy regressors – respectively mean predictor and conditional mean predictors (represented by the dotted horizontal lines), as structures are getting noisier. The intuition is that making meaningful predictions from (or close to) pure noise is difficult. Both types of models reach a similar accuracy for lower levels of noise. Initially, the error of model `C+L` is significantly larger than that of `L` as the model cannot guess what metal center the complex is going to feature. From Fig. 2b, we know that the metal center has a determinant influence on the binding energy.

3.4 Conditional generation and inverse-design

In this section, we analyze if OM-DIFF is able to generate novel optimized organometallic complexes that can effectively catalyze cross-coupling reactions.

Conditional generation We first investigate whether the conditioning mechanism works overall. To do so, we conditionally sample complexes whose target binding energies are spread across the property space. We study the generated compounds in terms of novelty, and examine their spread around the target value. We show that the conditioning mechanism is effective in parts of the space with sufficient data coverage, but that the effectiveness gradually decreases as we approach the tails of the distribution.

For each metal center, we choose the target values to be the following percentiles of the training data distribution: $[0.05, 0.25, 0.50, 0.75, 0.95]$. This is to make sure that the surrogate predictions remain somewhat reliable, as we cannot afford DFT calculations for all generated compounds. For each pair metal center - target value, we sample 10000 compounds, evaluate their binding energy using the surrogate model, and finally display the corresponding distribution.

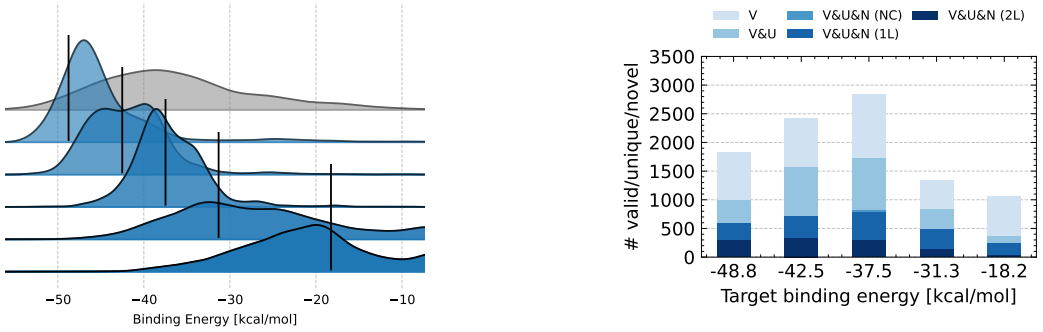


(a) (b)

Figure 7: **(a)** Residuals of the two variants of loss functions employed to train the screening regressor. 'MSE' refers to mean-square error, while 'revHuber' stands for reverse Huber. **(b)** Performance of the two variants of the time-conditioned regressor as a function of the diffusion time step. 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.

As an example in Fig. 8a, we show the conditional distributions obtained for Pd, while the same plots for the other metal centers are provided in Fig. S13. We can observe that for target values above the median, the conditional distributions tend to become more spread out. We hypothesize that this is due to the fact that the unconditional distribution, shown in grey in the back of Fig. 8a, is not symmetric around its median, and usually that target values above the median are more sparsely distributed. Nonetheless, we can effectively steer the conditional distribution. We also expect that the conditional distributions can be made sharper by upscaling the contribution of the guidance term in Eq. (16), at the cost of a lower validity and uniqueness.

As for property controllability, metrics such as validity, uniqueness and novelty are impacted by the target value as seen in Fig. 8b for Pd complexes, and in Fig. S14 for the other metal centers. For parts of the property space that are less well covered by training data, the different metrics tend to drop. Specifically, when targeting the 0.95 percentile, the conditional distribution tends to spread out, and (abnormally) high binding energies are predicted by the model. While the binding energy of these samples is probably mistakenly estimated by the surrogate, we observe that the conditional generative model is still able to produce novel complexes, namely around 250.



(a) (b)

Figure 8: **(a)** Binding energy distributions obtained through the conditional sampling of Pd, as evaluated by the surrogate. The distribution in grey in the background represents the training data distribution, i.e. DFT labels. Black vertical lines represent target values, and correspond to the $[0.05, 0.25, 0.50, 0.75, 0.95]$ percentiles of the training data distribution. Only valid samples were taken into account. **(b)** #Valid, #(Valid & Unique) and #(Valid & Unique & Novel) complexes for conditionally sampled Pd 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.

Inverse-designing optimized catalysts for the Suzuki reaction In the previous section, we showed that OM-DIFF could be effectively steered towards target binding energies of interest. Here, we attempt to design optimized catalyst that are relevant for the Suzuki cross-coupling reaction. Among the family of reactions under study, the Suzuki reaction has the narrowest plateau region of the volcano plot, spanning $[-32.1; -23.0]$ kcal/mol [58]. We therefore set the middle point of that interval, i.e. -27.55 kcal/mol, as a target value when performing conditional generation.

As previously, we generate 10000 complexes for Pd and Pt. After checking for validity and uniqueness, we only keep novel complexes. We additionally discard the ones with an estimated binding energy that does not fall within the range of interest. Of the remaining samples, we randomly keep 5 compounds for each metal center. For these compounds, we recompute the binding energies using DFT, employing the protocol described in Section 3.1. The novel catalysts are displayed in the two upper rows of Fig. 9, along with their binding energy estimated by the surrogate and calculated using DFT. For the Pd complexes, displayed in the top row, 4 out of 5 fall in the range of interest, and within 2 kcal/mol of the energy calculated with DFT. For the Pt complexes, all 5 complexes fall in the range of interest and within 2.5 kcal/mol of DFT. These slight discrepancies between DFT and surrogate predictions are in accordance with the errors reported Fig. S11, where the estimated MAE in that region of the property space is shown to be around $2-3$ kcal/mol.

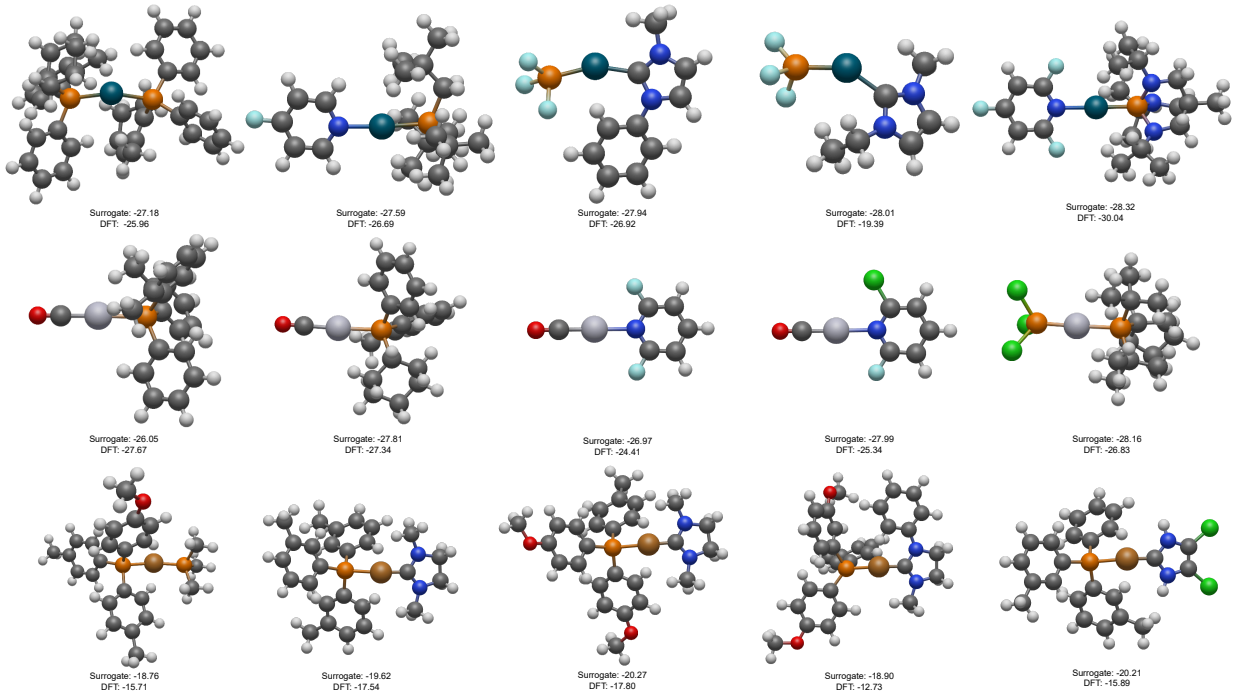


Figure 9: Overview of the novel complexes validated with DFT. (**Top row**) Pd complexes (**Middle row**) Pt complexes (**Bottom row**) Cu complexes. All binding energies are expressed in kcal/mol.

Inverse-designing optimized Cu catalysts We also tried to inverse-design Cu complexes relevant for the Suzuki reaction. This constitutes an interesting use-case as catalysts made of earth-abundant transition metals are highly desirable. As the screening surrogate could not identify valid samples in the said range of interest, we repeated the experiment with the 5% percentile, i.e. -20 kcal/mol, as a target value instead. Among the novel complexes generated, we kept 5 that were deemed close to the target value by the surrogate model. The considered complexes are illustrated in the bottom row of Fig. 9. With respect to the previous experiment, the spread between estimated and calculated binding energies appears to be larger. This can be explained by a less accurate surrogate model in that part of the property space, as illustrated in Fig. S11 where a sharp increase in MAE can be observed around -20 kcal/mol.

The inaccuracy of the screening surrogate is not the only explanation to the unsuccessful conditional generation for the initial target of -27.55 kcal/mol. The time-conditioned surrogate is also inaccurate in that region of the property space, and thereby likely to drive the generation process towards complexes with erroneous binding energies. Finally, as the generative model has only seen a handful of complexes in that part of the property space, we can imagine that it is not extremely good at modelling the distribution in that particular area.

4 Discussion and Conclusion

We have introduced OM-DIFF a framework for inverse-designing organometallic complexes for target applications. The framework is based on a guided equivariant denoising diffusion model specifically tailored for the generation of organometallic complexes with targeted properties. Instead of directly learning a conditional generative model, which will require large volume of computationally expensive data, this approach decouples the structure generative section from the conditional guidance towards coveted property, allowing the use of larger molecular databases for training and requiring only a limited set of task-specific labeled data for accurate model performance. To inverse design other organometallic catalysts for target reactions with OM-DIFF, one needs to establish (1) the chemical space of the complexes that catalyze the reaction and (2) the mechanism of the catalyzed reaction along with the rate-determining step.

We demonstrated the potential of the proposed framework on a dataset of catalyst candidates for a family of cross-coupling reactions. First, we showed that the increased expressivity of the denoising neural network combined with a proper modelling of the metal center enables effective unconditional generation of novel complexes. Second, we showed that the model offered controllability, and that sampling could effectively be steered across the property space while maintaining novelty in the sampled complexes. For the specific case of the Suzuki reaction, we further validated a handful of optimized complexes with DFT calculations. We could successfully generate promising Pd- and Pt-based complexes. The compounds were shown to have binding energy (for the activity determining step) within the range of interest. For Cu complexes, we could not generate complexes in the range of interest, highlighting the limitations of the proposed approach in parts of the property space sparsely represented in the training data. However, it is promising that when the target value was around a less extreme percentile of the property distribution, we could generate novel complexes featuring the prompted binding energy (or close to). This indicates that our framework could effectively be integrated in an active learning setup towards real discovery, where the training data is progressively extended towards properties of interest.

While we have demonstrated the applicability of OM-DIFF on the basis of generation of complexes from scratch, where the center is composed of one atom, the formulation introduced in Section 2.2 is more general, and naturally extends to problems where the context is composed of multiple atoms, for instance in cases where the catalyst is designed based on a handful of scaffolds [80], or in a functionalization setting [81].

We showed that goal-directed generation is attainable but that it remains a difficult endeavor with models trained offline with a static dataset. Other methods, based on GA or RL, are known to perform well in settings where they can query the function to be optimized. A hybrid method that uses an offline pretrained diffusion model, and that then further gets optimised online through RL or in an active learning setup is an interesting avenue for future. As samples with attractive properties often lie at the boundaries of the training data, such candidates can be evaluated with the QM method of choices, added to the training data, and the surrogate retrained on the augmented dataset. An iterative framework, where data is gathered continuously, is also a promising avenue that would allow the generative model to move towards regions of interest that were not well covered in the initial dataset.

We envision a significant scope for future work in both methodological development and application areas. While the framework was shown to work in a rather small data regime and with limited variety in the training data, pre-training on a relevant database [82], e.g. the TMQM database [16], before fine-tuning it in the chemical space of interest might allow for more valid, and novel molecules that lie beyond the neighborhood of the property-labeled training data. To be useful in practice, a generative model should suggest compounds that are valid and feasible to synthesize [8]. In OM-DIFF, the generation is steered towards promising parts of the chemical space based on a chemical target function, regardless of the potential validity and synthesizability. Within the guidance setup described in Section 2.3, the target function could further be modified to include the feedback of a classifier trained to distinguish between valid and invalid compounds, or the feedback of a surrogate trained to estimate synthesizability. Multi-property conditioning is also particularly relevant for performing inverse design of catalysts for more complex reactions where high catalytic activity requires optimal binding energy for multiple reaction steps.

Methodological improvement in OM-DIFF can include modelling of atom and bond types as categorical variables [83, 82] (instead of the continuous relaxation used in this work) and predicting the denoised structures directly, as it has been shown to work better for atomistic data [82]. Regarding the conditional sampling procedure, finding a better way to combine the feedback from the generative model and the regressor would be useful as well. As seen in Fig. S12, the time-conditioned is equivalent to a random guess on the early phases of the denoising process. If the surrogate provides uncertainty, this could also be leveraged to bias the generation, either to avoid uncertain regions or, in an active learning setting, to explore uncertain areas.

Code availability

Along with this paper, we release a code repository, that can be accessed at `https://github.com/frcnt/om-diff`, allowing other researchers to build upon our work.

Author Contributions

F.C., M.N.S. and A.B. conceived the research idea. F.C. developed the code with help from B.H., ran the experiments, analyzed the results, and wrote the original draft based on inputs from all authors. B.B. performed validation with DFT calculations. M.N.S. and A.B. supervised the research. All authors discussed, commented on, and revised the manuscript.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

The authors acknowledge financial support from the Independent Research Fund Denmark with project DELIGHT (Grant No. 0217-00326B).

References

[1] Jessica G Freeze, H Ray Kelly, and Victor S Batista. Search for catalysts by inverse design: artificial intelligence, mountain climbers, and alchemists. *Chemical reviews*, 119(11):6595–6612, 2019.

[2] Marco Foscato and Vidar R Jensen. Automated in silico design of homogeneous catalysts. *ACS catalysis*, 10(3):2354–2377, 2020.

[3] Benjamin Sanchez-Lengeling and Alán Aspuru-Guzik. Inverse molecular design using machine learning: Generative models for matter engineering. *Science*, 361(6400):360–365, 2018.

[4] Lowik Chanussot, Abhishek Das, Siddharth Goyal, Thibaut Lavril, Muhammed Shuaibi, Morgane Riviere, Kevin Tran, Javier Heras-Domingo, Caleb Ho, Weihua Hu, et al. Open catalyst 2020 (oc20) dataset and community challenges. *Acs Catalysis*, 11(10):6059–6072, 2021.

[5] Janice Lan, Aini Palizhati, Muhammed Shuaibi, Brandon M Wood, Brook Wander, Abhishek Das, Matt Uyttendaele, C Lawrence Zitnick, and Zachary W Ulissi. Adsorbml: a leap in efficiency for adsorption energy calculations using generalizable machine learning potentials. *npj Computational Materials*, 9(1):172, 2023.

[6] Dylan M Anstine and Olexandr Isayev. Generative models as an emerging paradigm in the chemical sciences. *Journal of the American Chemical Society*, 145(16):8736–8750, 2023.

[7] Seokho Kang and Kyunghyun Cho. Conditional molecular design with deep generative models. *Journal of chemical information and modeling*, 59(1):43–52, 2018.

[8] Wenhao Gao and Connor W Coley. The synthesizability of molecules proposed by generative models. *Journal of chemical information and modeling*, 60(12):5714–5723, 2020.

[9] Zhenpeng Yao, Benjamín Sánchez-Lengeling, N Scott Bobbitt, Benjamin J Bucior, Sai Govind Hari Kumar, Sean P Collins, Thomas Burns, Tom K Woo, Omar K Farha, Randall Q Snurr, et al. Inverse design of nanoporous crystalline reticular materials with deep generative models. *Nature Machine Intelligence*, 3(1):76–86, 2021.

[10] Xiang Fu, Tian Xie, Andrew Scott Rosen, Tommi S. Jaakkola, and Jake Allen Smith. MOFDiff: Coarse-grained diffusion for metal-organic framework design. In *The Twelfth International Conference on Learning Representations*, 2024.

[11] Arghya Bhowmik, Maitane Berecibar, Montse Casas-Cabanas, Gabor Csanyi, Robert Dominko, Kersti Hermansson, M Rosa Palacin, Helge S Stein, and Tejs Vegge. Implications of the battery 2030+ ai-assisted toolkit on future low-trl battery discoveries and chemistries. *Advanced Energy Materials*, 12(17):2102698, 2022.

[12] Raul Ortega Ochoa, Bardi Benediktsson, Renata Sechi, Peter Bjørn Jørgensen, and Arghya Bhowmik. Materials funnel 2.0–data-driven hierarchical search for exploration of vast chemical spaces. *Journal of Materials Chemistry A*, 11(48):26551–26561, 2023.

[13] Mingyang Wang, Chang-Yu Hsieh, Jike Wang, Dong Wang, Gaoqi Weng, Chao Shen, Xiaojun Yao, Zhitong Bing, Honglin Li, Dongsheng Cao, et al. Relation: A deep generative model for structure-based de novo drug design. *Journal of Medicinal Chemistry*, 65(13):9478–9492, 2022.

[14] John J Irwin and Brian K Shoichet. Zinc- a free database of commercially available compounds for virtual screening. *Journal of chemical information and modeling*, 45(1):177–182, 2005.

[15] Maho Nakata and Tomomi Shimazaki. Pubchemqc project: a large-scale first-principles electronic structure database for data-driven chemistry. *Journal of chemical information and modeling*, 57(6):1300–1308, 2017.

[16] David Balcells and Bastian Bjerkem Skjelstad. tmqm dataset—quantum geometries and properties of 86k transition metal complexes. *Journal of chemical information and modeling*, 60(12):6135–6146, 2020.

[17] Nicola De Cao and Thomas Kipf. Molgan: An implicit generative model for small molecular graphs. *arXiv preprint arXiv:1805.11973*, 2018.

[18] Jiaxuan You, Rex Ying, Xiang Ren, William Hamilton, and Jure Leskovec. Graphrnn: Generating realistic graphs with deep auto-regressive models. In *International conference on machine learning*, pages 5708–5717. PMLR, 2018.

[19] Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Junction tree variational autoencoder for molecular graph generation. In *International conference on machine learning*, pages 2323–2332. PMLR, 2018.

[20] Marco Foscato, Giovanni Occhipinti, Vishwesh Venkatraman, Bjørn K Alsberg, and Vidar R Jensen. Automated design of realistic organometallic molecules from fragments. *Journal of Chemical Information and Modeling*, 54(3):767–780, 2014.

[21] Jan H Jensen. A graph-based genetic algorithm and generative model/monte carlo tree search for the exploration of chemical space. *Chemical science*, 10(12):3567–3572, 2019.

[22] Julius Seumer, Jonathan Kirschner Solberg Hansen, Mogens Brøndsted Nielsen, and Jan H Jensen. Computational evolution of new catalysts for the morita–baylis–hillman reaction. *Angewandte Chemie International Edition*, 62(18):e202218565, 2023.

[23] Magnus Strandgaard, Julius Seumer, Bardi Benediktsson, Arghya Bhowmik, Tejs Vegge, and Jan H Jensen. Genetic algorithm-based re-optimization of the schrock catalyst for dinitrogen fixation. *PeerJ physical chemistry*, 5:e30, 2023.

[24] Oliver Schilter, Alain Vaucher, Philippe Schwaller, and Teodoro Laino. Designing catalysts with deep generative models and computational data. a case study for suzuki cross coupling reactions. *Digital Discovery*, 2(3):728–735, 2023.

[25] Niklas Gebauer, Michael Gastegger, and Kristof Schütt. Symmetry-adapted generation of 3d point sets for the targeted discovery of molecules. *Advances in neural information processing systems*, 32:7566—-7578, 2019.

[26] Emiel Hoogeboom, Vıctor Garcia Satorras, Clément Vignac, and Max Welling. Equivariant diffusion for molecule generation in 3d. In *International conference on machine learning*, pages 8867–8887. PMLR, 2022.

[27] Niklas WA Gebauer, Michael Gastegger, Stefaan SP Hessmann, Klaus-Robert Müller, and Kristof T Schütt. Inverse design of 3d molecular structures with conditional generative neural networks. *Nature communications*, 13(1):973, 2022.

[28] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015.

[29] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[30] Minkai Xu, Lantao Yu, Yang Song, Chence Shi, Stefano Ermon, and Jian Tang. Geodiff: A geometric diffusion model for molecular conformation generation. In *International Conference on Learning Representations*, 2021.

[31] Ilia Igashov, Hannes Stärk, Clément Vignac, Victor Garcia Satorras, Pascal Frossard, Max Welling, Michael Bronstein, and Bruno Correia. Equivariant 3d-conditional diffusion models for molecular linker design. *arXiv preprint arXiv:2210.05274*, 2022.

[32] Arne Schneuing, Yuanqi Du, Charles Harris, Arian Jamasb, Ilia Igashov, Weitao Du, Tom Blundell, Pietro Lió, Carla Gomes, Max Welling, et al. Structure-based drug design with equivariant diffusion models. *arXiv preprint arXiv:2210.13695*, 2022.

[33] Jiaqi Guan, Wesley Wei Qian, Xingang Peng, Yufeng Su, Jian Peng, and Jianzhu Ma. 3d equivariant diffusion for target-aware molecule generation and affinity prediction. In *The Eleventh International Conference on Learning Representations*, 2022.

[34] Fan Bao, Min Zhao, Zhongkai Hao, Peiyao Li, Chongxuan Li, and Jun Zhu. Equivariant energy-guided sde for inverse molecular design. In *The eleventh international conference on learning representations*, 2022.

[35] Tomer Weiss, Eduardo Mayo Yanes, Sabyasachi Chakraborty, Luca Cosmo, Alex M Bronstein, and Renana Gershoni-Poranne. Guided diffusion for inverse molecular design. *Nature Computational Science*, 3(10):873–882, 2023.

[36] Diederik Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. *Advances in neural information processing systems*, 34:21696–21707, 2021.

[37] Kristof Schütt, Oliver Unke, and Michael Gastegger. Equivariant message passing for the prediction of tensorial properties and molecular spectra. In *International Conference on Machine Learning*, pages 9377–9388. PMLR, 2021.

[38] Vıctor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E (n) equivariant graph neural networks. In *International conference on machine learning*, pages 9323–9332. PMLR, 2021.

[39] Chaitanya K. Joshi, Cristian Bodnar, Simon V Mathis, Taco Cohen, and Pietro Lio. On the expressive power of geometric graph neural networks, 2023.

[40] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

[41] Jiaming Song, Qinsheng Zhang, Hongxu Yin, Morteza Mardani, Ming-Yu Liu, Jan Kautz, Yongxin Chen, and Arash Vahdat. Loss-guided diffusion models for plug-and-play controllable generation. In *International Conference on Machine Learning*, pages 32483–32498. PMLR, 2023.

[42] Pingyu Zhang and Peter J Sadler. Advances in the design of organometallic anticancer complexes. *Journal of Organometallic Chemistry*, 839:5–14, 2017.

[43] Kyle A Williams, Andrew J Boydston, and Christopher W Bielawski. Main-chain organometallic polymers: synthetic strategies, applications, and perspectives. *Chemical Society Reviews*, 36(5):729–744, 2007.

[44] Jan Kalinowski, Valeria Fattori, Massimo Cocchi, and JA Gareth Williams. Light-emitting devices based on organometallic platinum complexes as emitters. *Coordination Chemistry Reviews*, 255(21-22):2401–2425, 2011.

[45] Wai-Yeung Wong and Cheuk-Lam Ho. Organometallic photovoltaics: a new and versatile approach for harvesting solar energy using conjugated polymetallaynes. *Accounts of chemical research*, 43(9):1246–1256, 2010.

[46] Dan-Yang Wang, Ruilan Liu, Wei Guo, Gang Li, and Yongzhu Fu. Recent advances of organometallic complexes for rechargeable batteries. *Coordination Chemistry Reviews*, 429:213650, 2021.

[47] Jonathan W Steed. Coordination and organometallic compounds as anion receptors and sensors. *Chemical Society Reviews*, 38(2):506–519, 2009.

[48] Matthew D Jones, Robert Raja, John Meurig Thomas, Brian FG Johnson, Dewi W Lewis, Jacques Rouzaud, and Kenneth DM Harris. Enhancing the enantioselectivity of novel homogeneous organometallic hydrogenation catalysts. *Angewandte Chemie International Edition*, 42(36):4326–4331, 2003.

[49] Jola Pospech, Ivana Fleischer, Robert Franke, Stefan Buchholz, and Matthias Beller. Alternative metals for homogeneous catalyzed hydroformylation reactions. *Angewandte Chemie International Edition*, 52(10):2852–2872, 2013.

[50] Tina M Trnka and Robert H Grubbs. The development of l2x2ru chr olefin metathesis catalysts: an organometallic success story. *Accounts of Chemical Research*, 34(1):18–29, 2001.

[51] Zhenxing Xi, Xiaoming Zhang, Wanzhi Chen, Shizhou Fu, and Daqi Wang. Synthesis and structural characterization of nickel (ii) complexes supported by pyridine-functionalized n-heterocyclic carbene ligands and their catalytic acitivities for suzuki coupling. *Organometallics*, 26(26):6636–6642, 2007.

[52] Arun Kumar, Gyandshwar Kumar Rao, Satyendra Kumar, and Ajai K Singh. Formation and role of palladium chalcogenide and other species in suzuki–miyaura and heck c–c coupling reactions catalyzed with palladium (ii) complexes of organochalcogen ligands: realities and speculations. *Organometallics*, 33(12):2921–2943, 2014.

[53] Pablo Espinet and Antonio M Echavarren. The mechanisms of the stille reaction. *Angewandte Chemie International Edition*, 43(36):4704–4734, 2004.

[54] Akira Suzuki. Cross-coupling reactions of organoboranes: an easy way to construct c–c bonds (nobel lecture). *Angewandte Chemie International Edition*, 30(50):6722–6737, 2011.

[55] Norio Miyaura and Akira Suzuki. Palladium-catalyzed cross-coupling reactions of organoboron compounds. *Chemical reviews*, 95(7):2457–2483, 1995.

[56] Michael Busch, Matthew D Wodrich, and Clémence Corminboeuf. A generalized picture of c–c cross-coupling. *ACS Catalysis*, 7(9):5643–5653, 2017.

[57] Boodsarin Sawatlon, Matthew D Wodrich, Benjamin Meyer, Alberto Fabrizio, and Clémence Corminboeuf. Data mining the c- c cross-coupling genome. *ChemCatChem*, 11(16):4096–4107, 2019.

[58] Benjamin Meyer, Boodsarin Sawatlon, Stefan Heinen, O Anatole Von Lilienfeld, and Clémence Corminboeuf. Machine learning meets volcano plots: computational discovery of cross-coupling catalysts. *Chemical science*, 9(35):7069–7077, 2018.

[59] Thomas A Halgren. Merck molecular force field. i. basis, form, scope, parameterization, and performance of mmff94. *Journal of computational chemistry*, 17(5-6):490–519, 1996.

[60] Frank Neese. The orca program system. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 2(1):73–78, 2012.

[61] Frank Neese. Software update: The orca program system—version 5.0. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 12(5):e1606, 2022.

[62] Axel D Beck. Density-functional thermochemistry. iii. the role of exact exchange. *J. Chem. Phys*, 98(7):5648–6, 1993.

[63] Axel D Becke. Density-functional exchange-energy approximation with correct asymptotic behavior. *Physical review A*, 38(6):3098, 1988.

[64] Chengteh Lee, Weitao Yang, and Robert G Parr. Development of the colle-salvetti correlation-energy formula into a functional of the electron density. *Physical review B*, 37(2):785, 1988.

[65] J Stephen Binkley, John A Pople, and Warren J Hehre. Self-consistent molecular orbital methods. 21. small split-valence basis sets for first-row elements. *Journal of the American Chemical Society*, 102(3):939–947, 1980.

[66] William J Pietro, Michelle M Francl, Warren J Hehre, Douglas J DeFrees, John A Pople, and J Stephen Binkley. Self-consistent molecular orbital methods. 24. supplemented small split-valence basis sets for second-row elements. *Journal of the American Chemical Society*, 104(19):5039–5048, 1982.

[67] KD Dobbs and WJ Hehre. Molecular orbital theory of the properties of inorganic and organometallic compounds 5. extended basis sets for first-row transition metals. *Journal of Computational Chemistry*, 8(6):861–879, 1987.

[68] KD Dobbs and WJ Hehre. Molecular orbital theory of the properties of inorganic and organometallic compounds. 6. extended basis sets for second-row transition metals. *Journal of computational chemistry*, 8(6):880–893, 1987.

[69] Florian Weigend and Reinhart Ahlrichs. Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for h to rn: Design and assessment of accuracy. *Physical Chemistry Chemical Physics*, 7(18):3297–3305, 2005.

[70] Benjamin P Pritchard, Doaa Altarawy, Brett Didier, Tara D Gibson, and Theresa L Windus. New basis set exchange: An open, up-to-date resource for the molecular sciences community. *Journal of chemical information and modeling*, 59(11):4814–4820, 2019.

[71] Stefan Grimme, Jens Antony, Stephan Ehrlich, and Helge Krieg. A consistent and accurate ab initio parametrization of density functional dispersion correction (dft-d) for the 94 elements h-pu. *The Journal of chemical physics*, 132(15), 2010.

[72] Frank Neese, Frank Wennmohs, Andreas Hansen, and Ute Becker. Efficient, approximate and parallel hartree–fock and hybrid dft calculations. a 'chain-of-spheres' algorithm for the hartree–fock exchange. *Chemical Physics*, 356(1-3):98–109, 2009.

[73] Róbert Izsák and Frank Neese. An overlap fitted chain of spheres exchange method. *The Journal of chemical physics*, 135(14), 2011.

[74] Georgi L Stoychev, Alexander A Auer, and Frank Neese. Automatic generation of auxiliary basis sets. *Journal of chemical theory and computation*, 13(2):554–562, 2017.

[75] Florian Weigend. Accurate coulomb-fitting basis sets for h to rn. *Physical chemistry chemical physics*, 8(9):1057–1065, 2006.

[76] Julius Seumer and Jan H Jensen. Beyond predefined ligand libraries: A genetic algorithm approach for de novo discovery of catalysts for the suzuki coupling reactions. 2024.

[77] Christoph Bannwarth, Sebastian Ehlert, and Stefan Grimme. Gfn2-xtb—an accurate and broadly parametrized self-consistent tight-binding quantum chemical method with multipole electrostatics and density-dependent dispersion contributions. *Journal of chemical theory and computation*, 15(3):1652–1671, 2019.

[78] Alex Morehead and Jianlin Cheng. Geometry-complete diffusion for 3d molecule generation. In *ICLR 2023 - Machine Learning for Drug Discovery workshop*, 2023.

[79] Greg Landrum et al. Rdkit: A software suite for cheminformatics, computational chemistry, and predictive modeling. *Greg Landrum*, 8:31, 2013.

[80] Annika M Krieger, Vivek Sinha, Adarsh V Kalikadien, and Evgeny A Pidko. Metal-ligand cooperative activation of hx (x= h, br, or) bond on mn based pincer complexes. *Zeitschrift für anorganische und allgemeine Chemie*, 647(14):1486–1494, 2021.

[81] Adarsh V Kalikadien, Evgeny A Pidko, and Vivek Sinha. Chemspax: exploration of chemical space by automated functionalization of molecular scaffold. *Digital Discovery*, 1(1):8–25, 2022.

[82] Tuan Le, Julian Cremer, Frank Noe, Djork-Arné Clevert, and Kristof T Schütt. Navigating the design space of equivariant diffusion-based generative models for de novo 3d molecule generation. In *The Twelfth International Conference on Learning Representations*, 2024.

[83] Clement Vignac, Nagham Osman, Laura Toni, and Pascal Frossard. Midi: Mixed graph and 3d denoising diffusion for molecule generation. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 560–576. Springer, 2023.

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 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 [26], based on EGNN [38].

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 [37]. 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 [37] 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.

S1.2 Denoiser performance

In Fig. S1, we display the loss of the different variants of OM-DIFF. The loss is evaluated on a held-out validation 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 $\mathcal{C}^{(L)} = \frac{1}{\alpha_t}(\mathcal{C}_t^{(L)} - \sigma_t \epsilon)$.

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 [26]

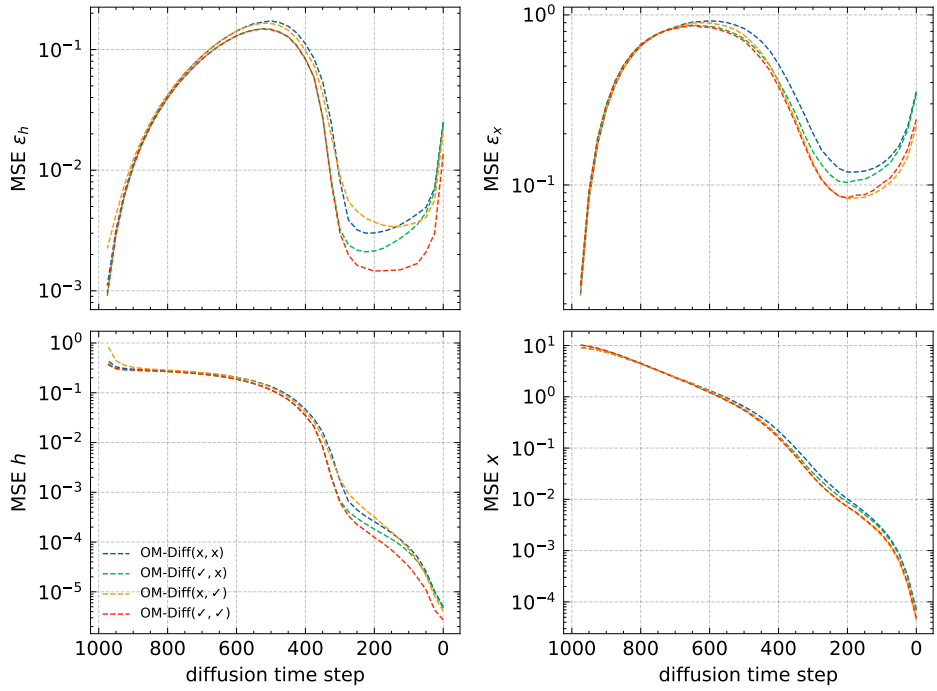


Figure S1: Evaluation of the different denoisers. **(Top row)** Prediction error for the different denoisers. **(Bottom row)** Resulting error on the estimated denoised sample for each denoiser.

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);

3. **(RDKit check)** The ligands, i.e. complex where the TM has been removed, have to be valid according to RDKit [79]. 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.

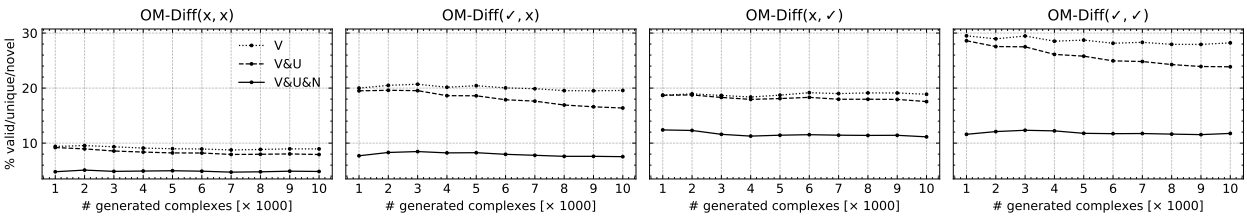


Figure S2: %Valid, %(Valid & Unique) and %(Valid & Unique & Novel) complexes for each variant of the generative model, as a function of the number of generated complexes. Novelty can come from novel combinations of existing compounds or from novel ligands.

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

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

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

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

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

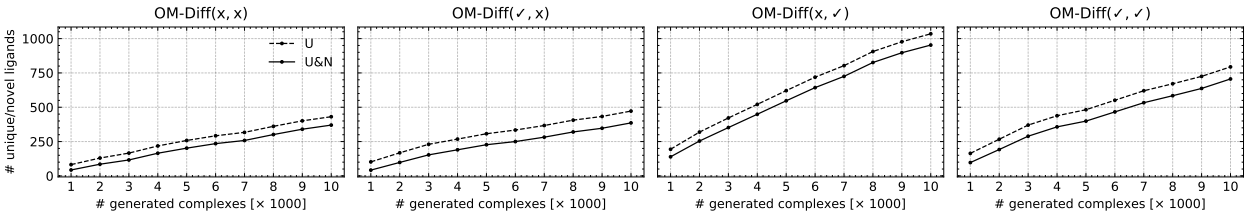


Figure S3: #Unique and #(Unique & Novel) ligands for each variant of the generative model, as a function of the number of generated complexes.

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 $\text{L}_{1,2} - \text{M}$ distances and the $\text{L}_1 - \text{M} - \text{L}_2$ angle. We provide all numerical results in Table S3, along with the corresponding distribution in Fig. S4, and the details of $\text{L}_{1,2}$-M for each metal center Fig. S5.

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,

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

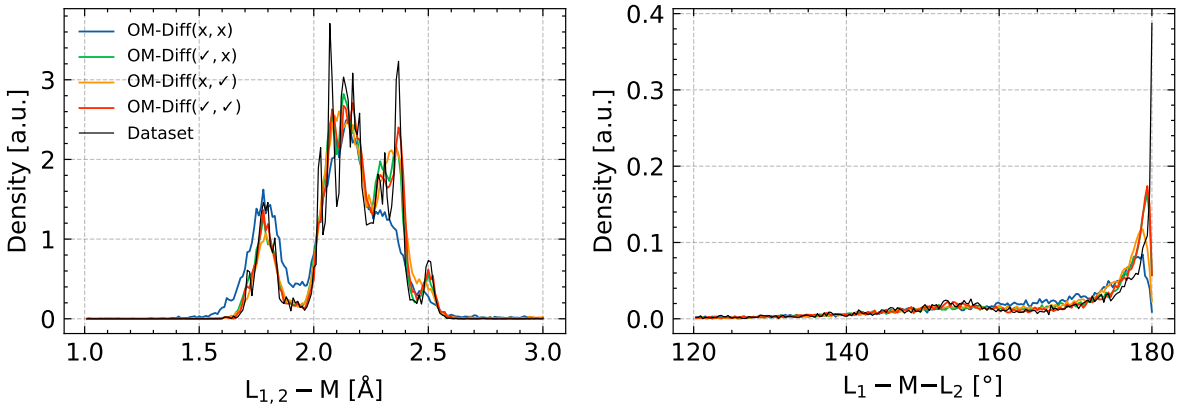


Figure S4: **(Top)** Distribution of $L_{1,2}$-M distances. **(Bottom)** Distribution of L_1-M-L_2 angles.

we additionally provide the cumulative distribution of predictive uncertainty, estimated as the standard deviation across the same ensemble, detailed for each metal center.

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

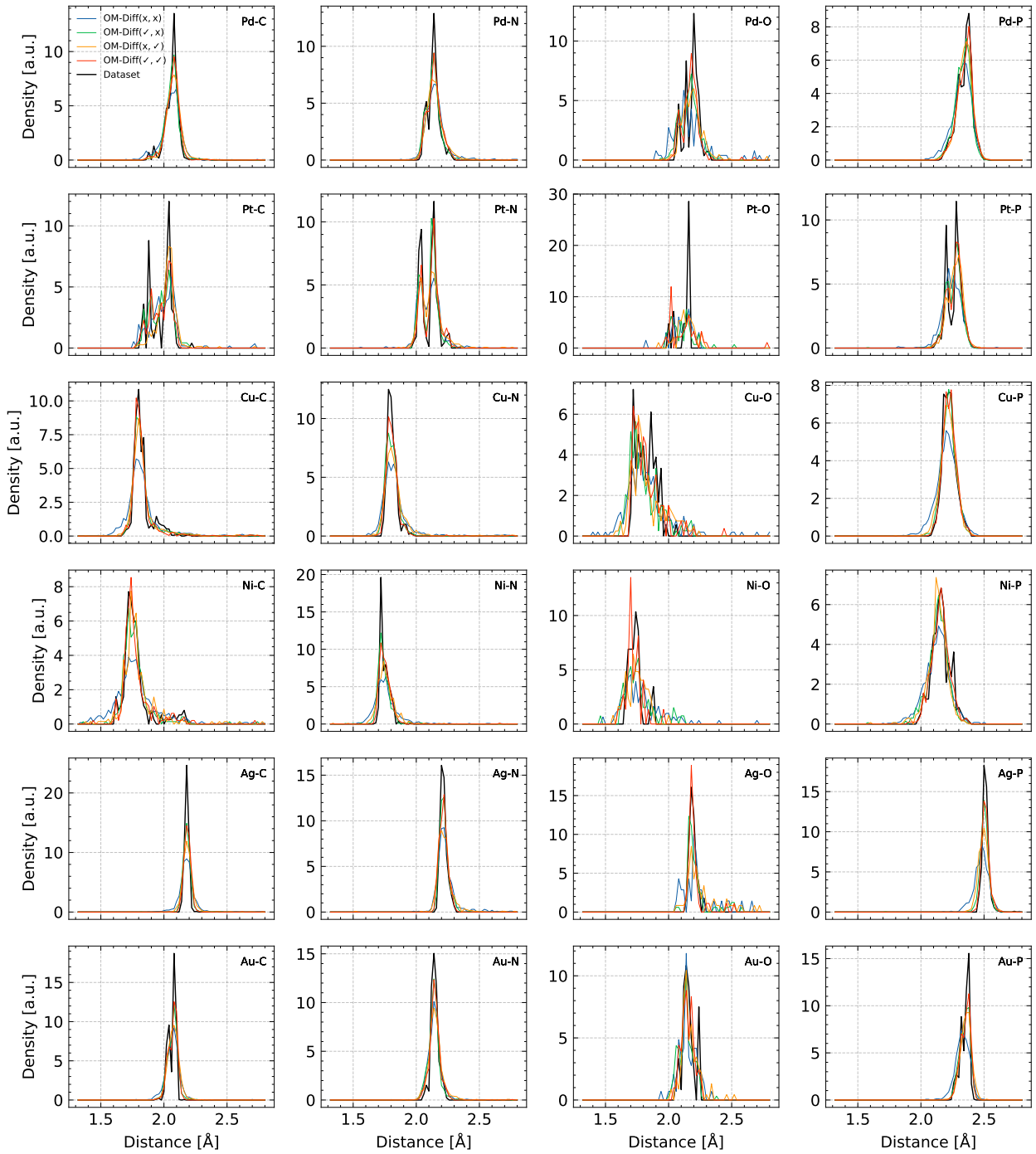


Figure S5: Distribution of bonds involving the metal center for the different variants of the diffusion generative model.

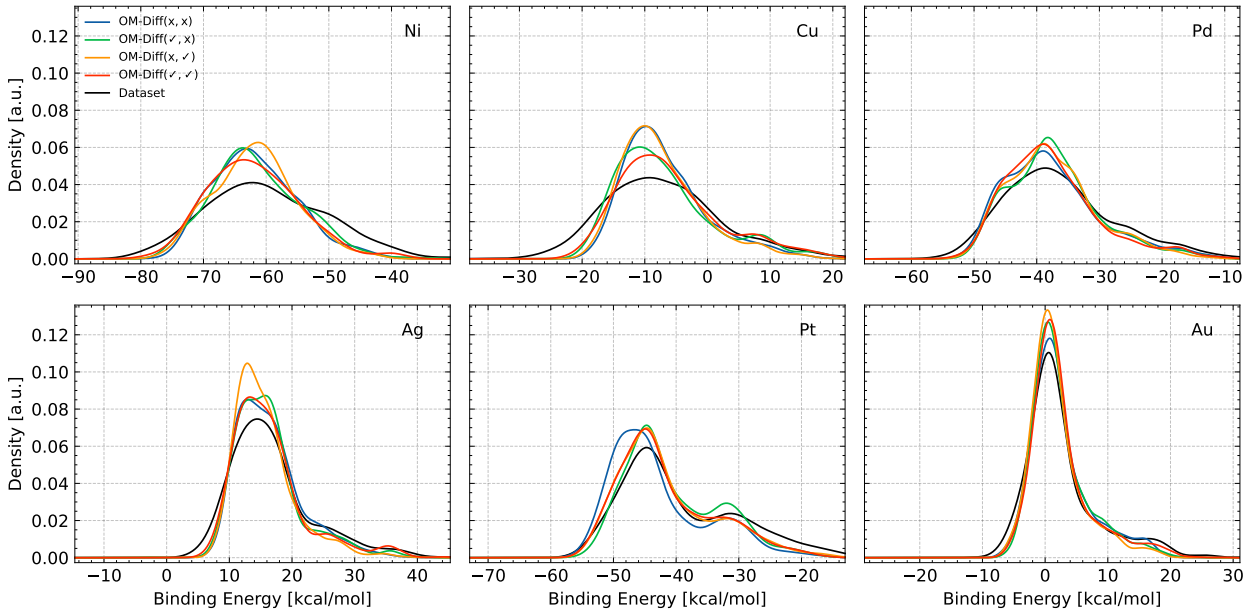


Figure S6: Distribution of binding energies for unconditionally generated samples.

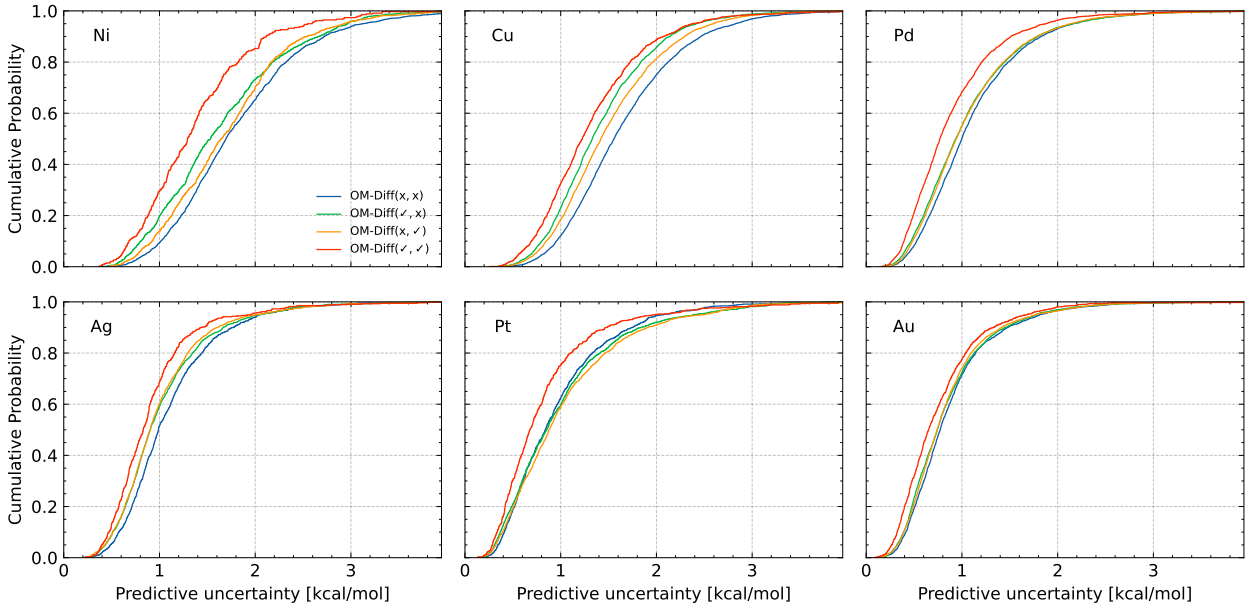


Figure S7: Predictive uncertainty, as estimated by an ensemble of 10 surrogates, evaluated on the samples generated by the different variants of the generative diffusion model. Similar to Fig. 6a, with each metal centre illustrated separately.

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

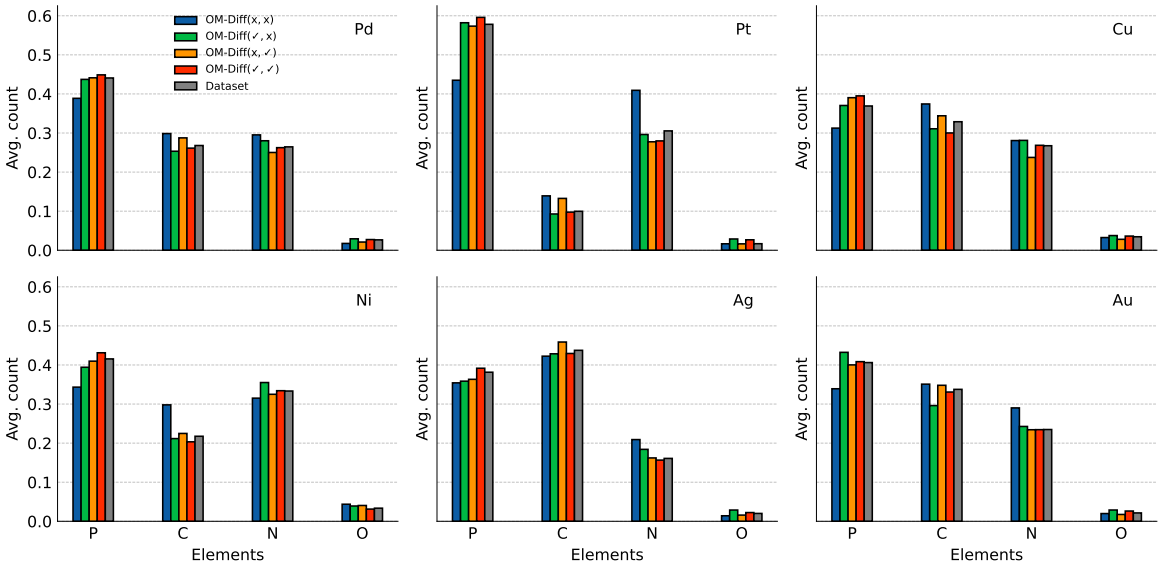


Figure S8: Distribution of proximal atoms, in unconditional sampling. Proximal atoms are defined as the atoms of the ligands to which the metal center is bound.

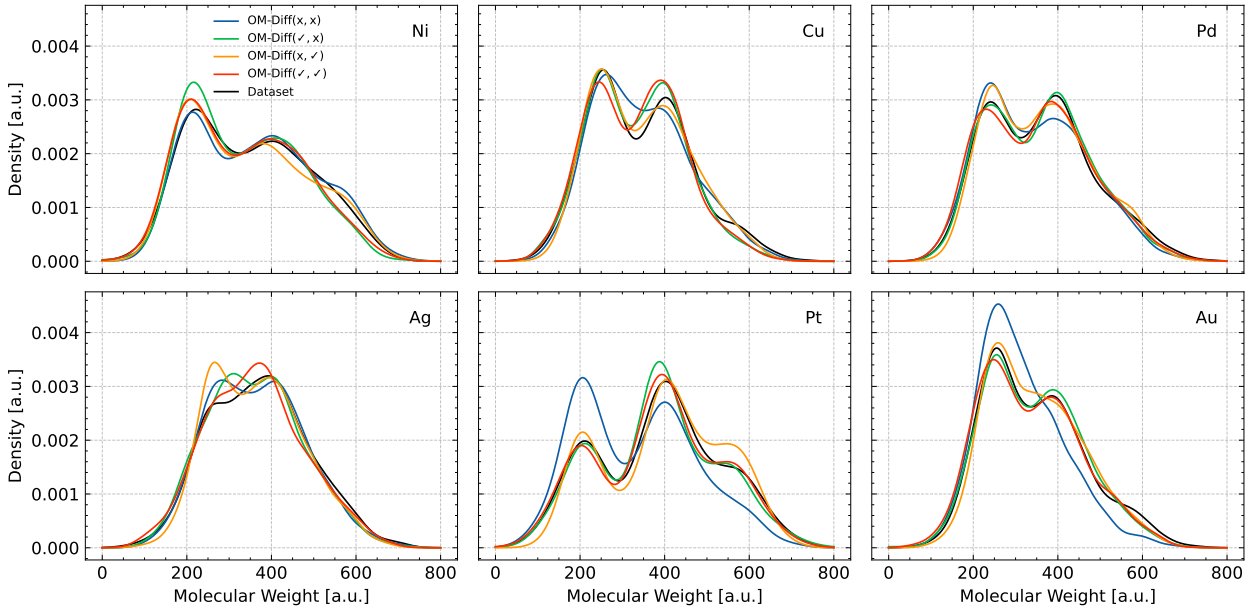


Figure S9: Distribution of molecular weight of the samples generated by the different variants of the generative diffusion model, detailed for each model. The mass of the metal center has been subtracted such that only the weight of the ligands is displayed.

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 [58] and string+MLP [24], we report results are from the respective papers.

	μ_M	SLATM [58]	string+MLP [24]	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, while

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.

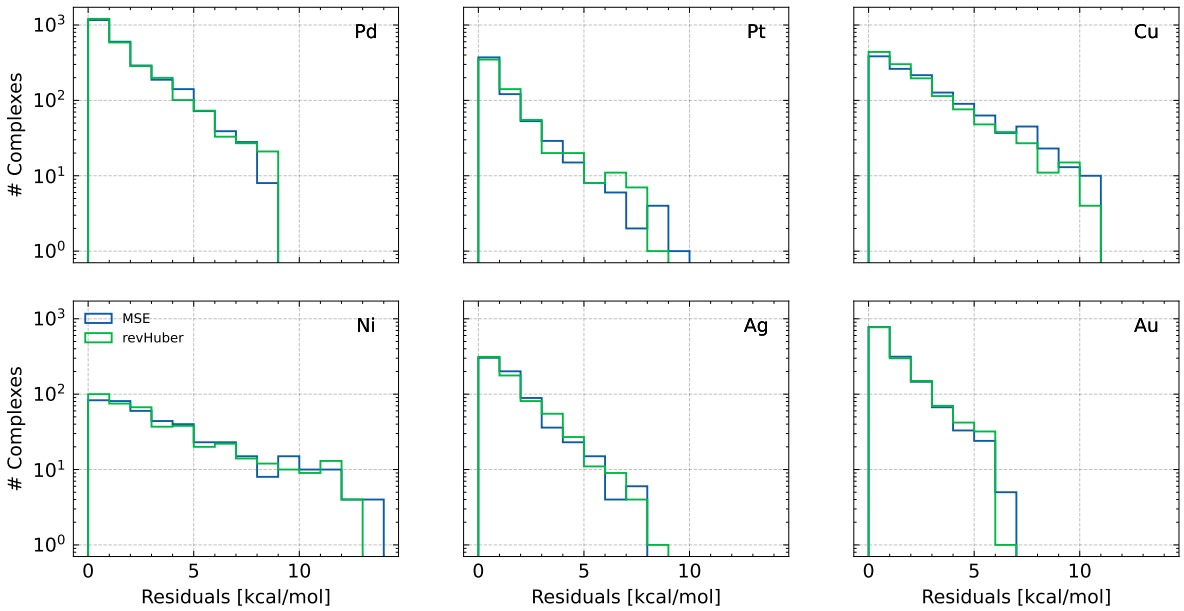


Figure S10: Residuals of the two variants of loss functions employed to train the screening regressor for each metal center. 'MSE' refers to mean-square error, while 'revHuber' stands for reverse Huber.

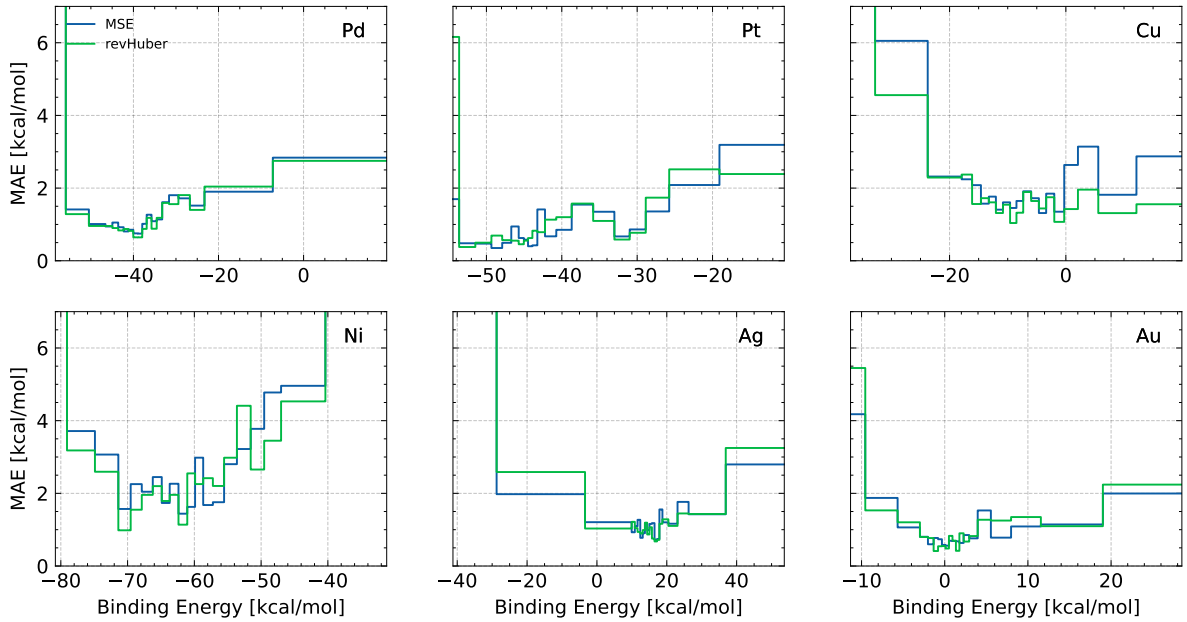


Figure S11: MAE across the property space of the two variants of loss functions employed to train the screening regressor, shown for each metal center. 'MSE' refers to mean-square error, while 'revHuber' stands for reverse Huber.

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 [24], we report results from the paper.

	μ_M	string+MLP [24]	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 [24], we report results are from the respective papers.

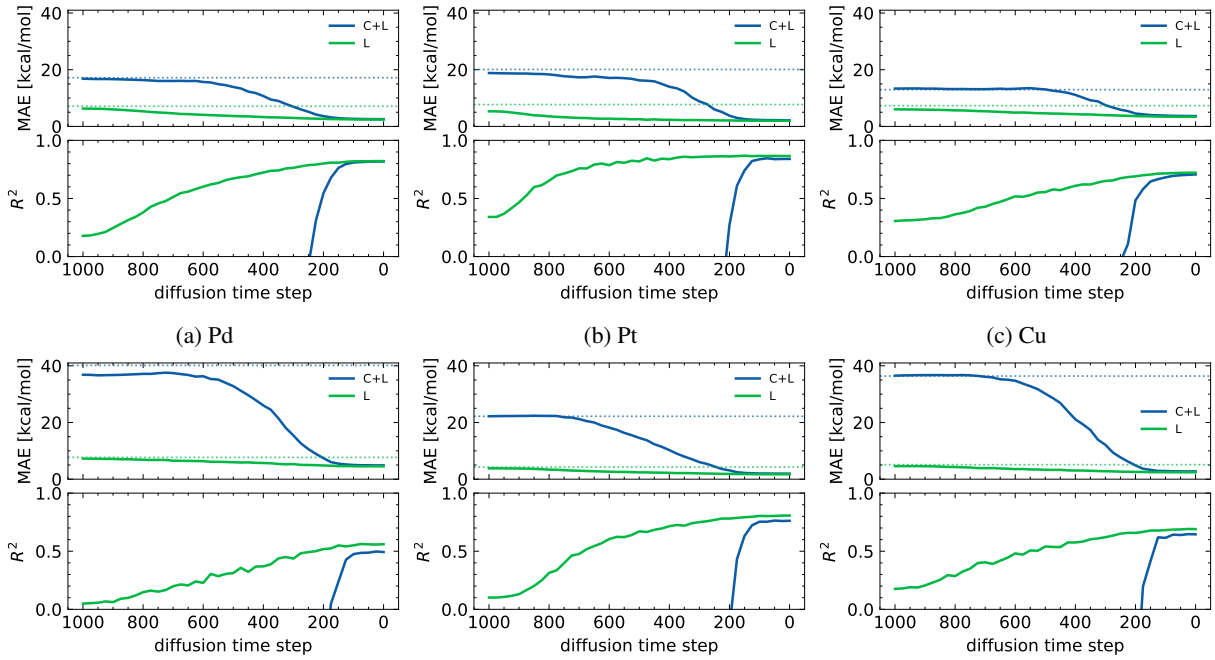
	μ_M	string+MLP [24]	Ours (MSE)	Ours (revHuber)
All	41.71 ± 12.38	26.02	32.09 ± 16.94	32.36 ± 16.37
Ni	23.79 ± 5.09	–	17.03 ± 4.13	17.33 ± 6.65
Cu	27.56 ± 2.96	–	16.90 ± 6.74	17.49 ± 6.50
Pd	35.84 ± 12.64	–	21.83 ± 15.97	21.45 ± 15.33
Ag	26.21 ± 13.90	–	18.77 ± 15.48	17.96 ± 14.88
Pt	24.25 ± 4.14	–	12.35 ± 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 [24], we report results are from the respective papers.

	string+MLP [24]	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.



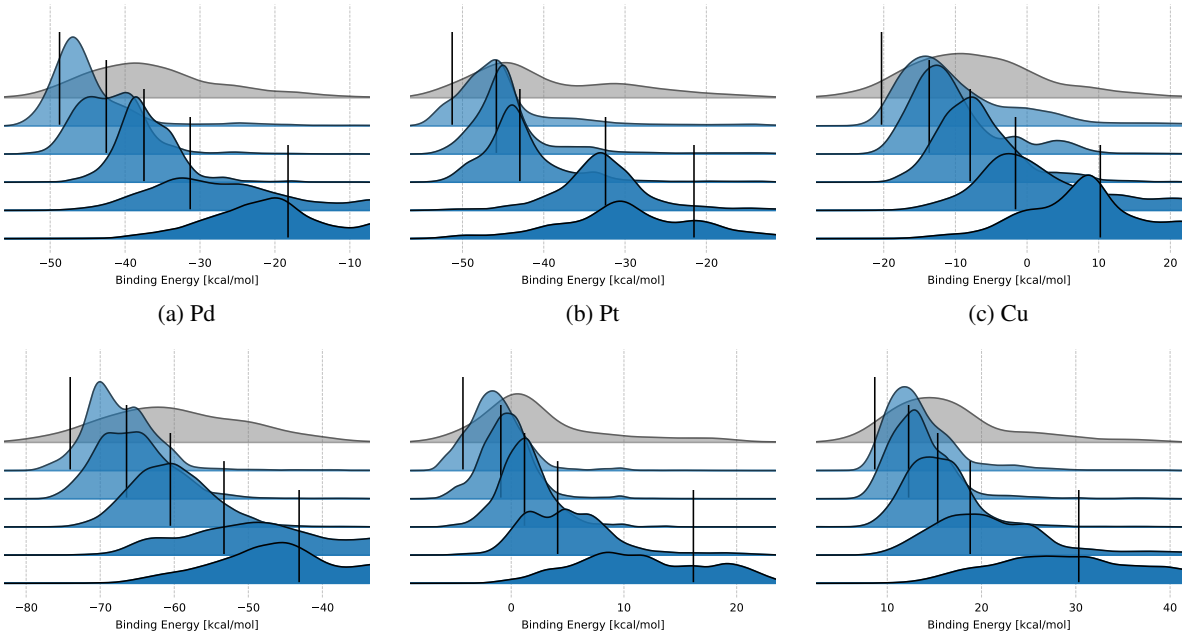
(a) Pd (b) Pt (c) Cu

(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

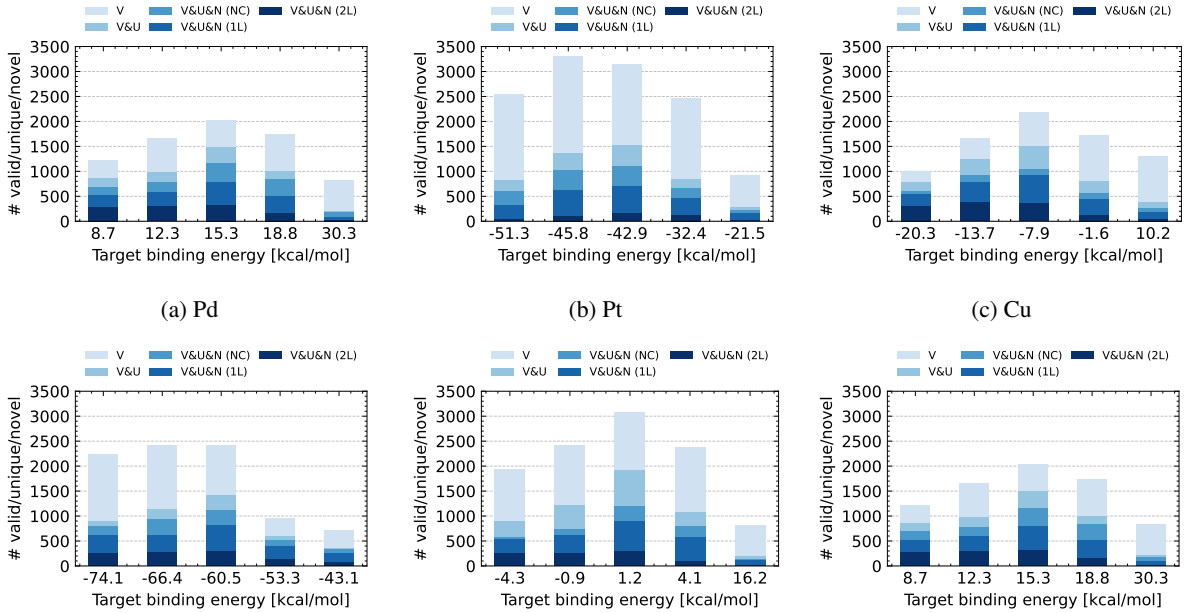
In this section, we provide the conditional distributions for metal center in Fig. S13, and the corresponding Validity/Uniqueness/Novelty breakdowns in Fig. S14.



(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.

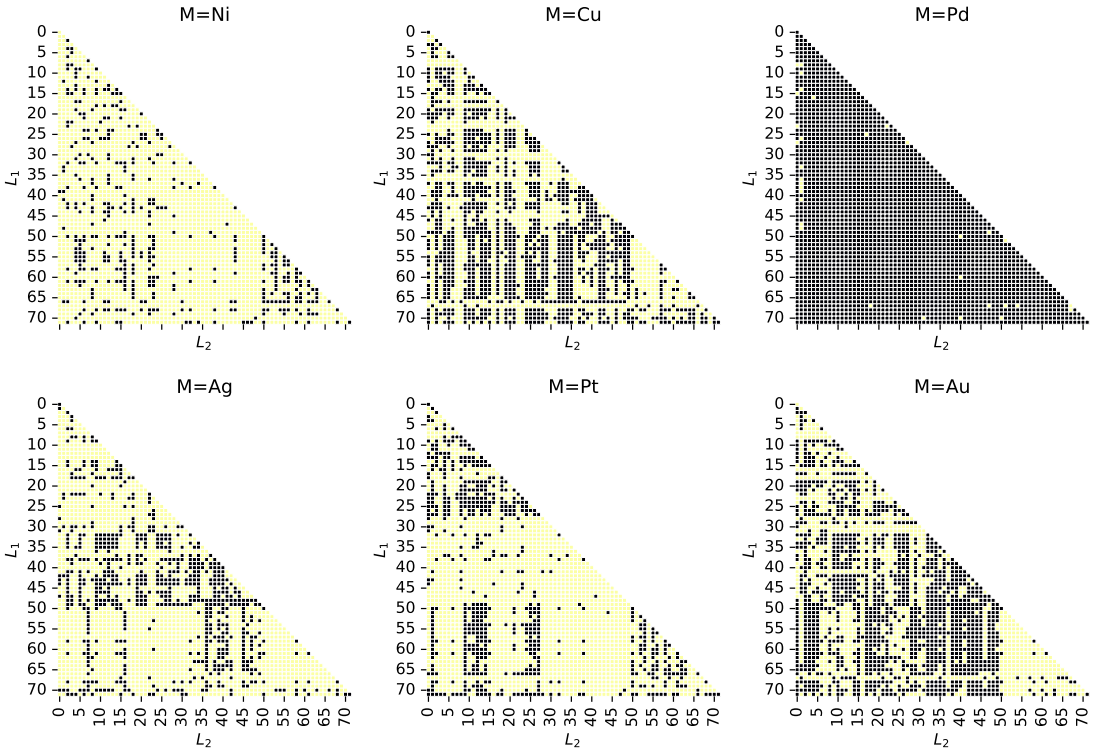


Figure S15: Combinations metal-ligand $L_1 - M - L_2$ composing the dataset. Black squares represent data points present in the dataset.

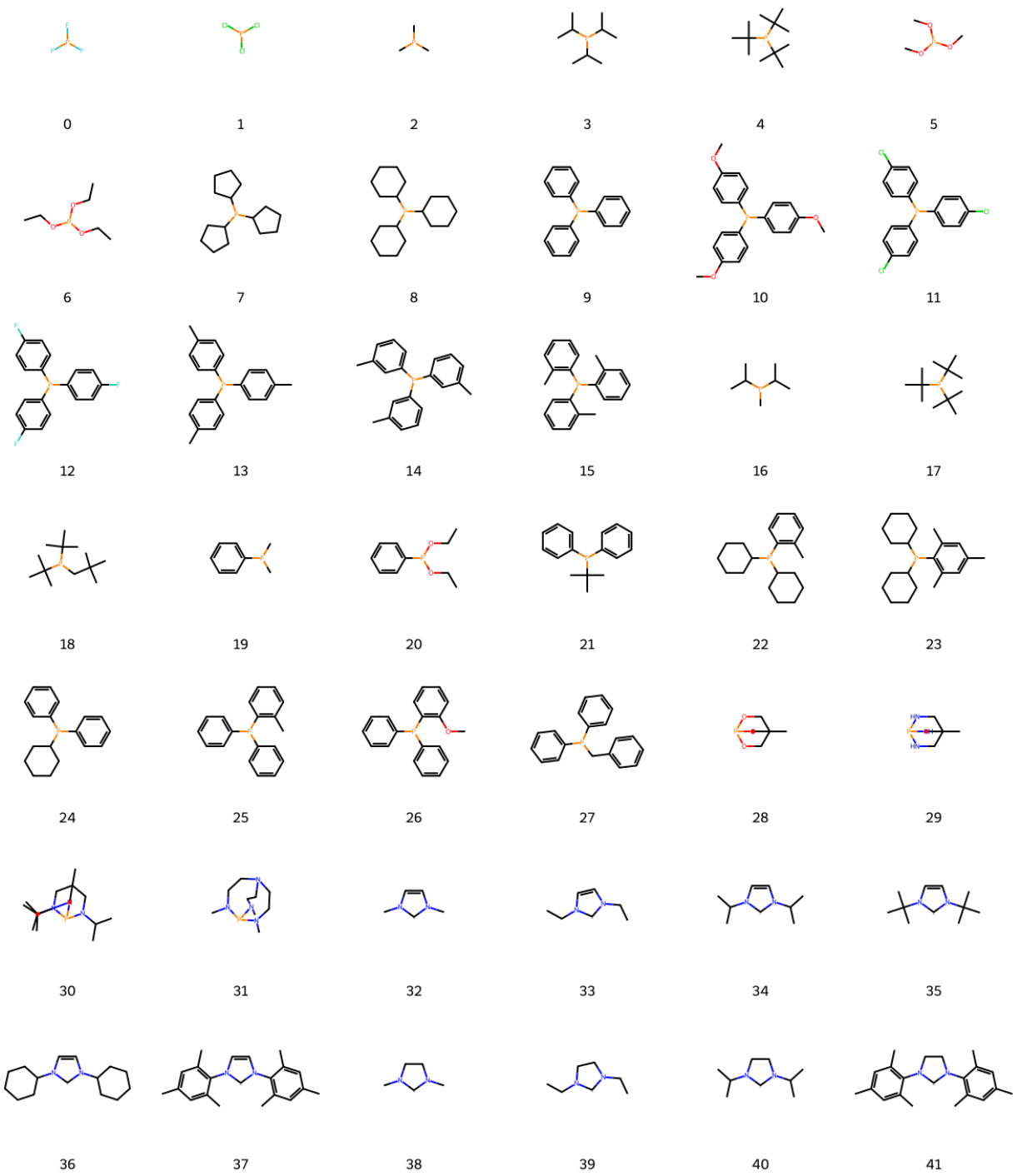


Figure S16: Ligands 0-41 used to build the dataset[58].

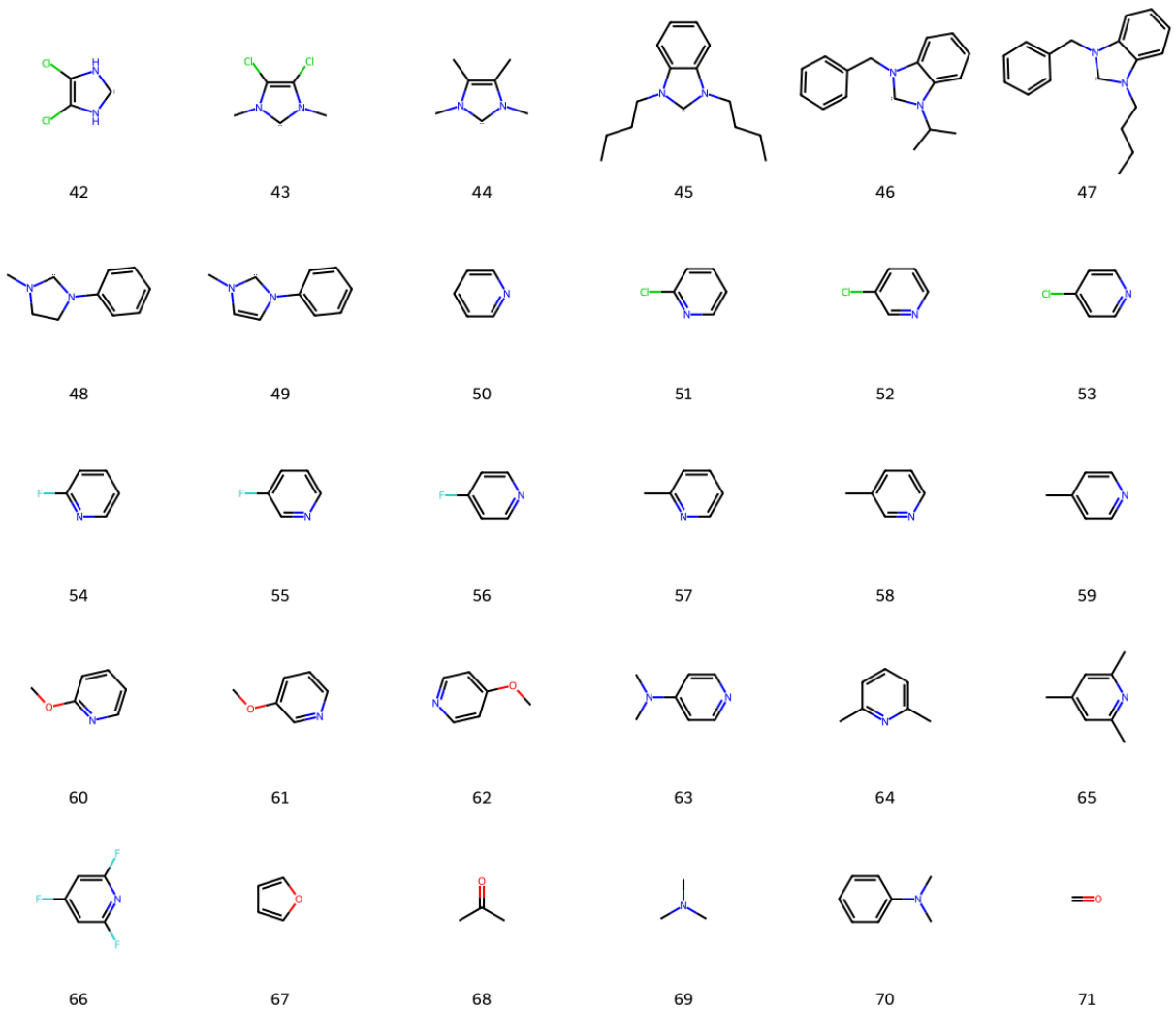


Figure S17: Ligands 42-72 used to build the dataset [58].