IOP Publishing Mach. Learn.: Sci. Technol. 6 (2025) 025057 https://doi.org/10.1088/2632-2153/addc32

MACHINE LEARNING
Science and Technology

PAPER

Improving generative inverse design of molecular catalysts in small data regime

François Cornet[1,2], Pratham Deshmukh[1], Bardi Benediktsson[1], Mikkel N Schmidt[2] and Arghya Bhowmik[1,*]

[1] Department of Energy Conversion and Storage, Technical University of Denmark, Kgs. Lyngby, Denmark
[2] Department of Applied Mathematics and Computer Science, Technical University of Denmark, Kgs. Lyngby, Denmark
[*] Author to whom any correspondence should be addressed.

E-mail: arbh@dtu.dk

Keywords: inverse design, multi-task learning, data representation, transition metal complex, diffusion model, catalyst design, small data ML

OPEN ACCESS

RECEIVED
10 January 2025

REVISED
24 April 2025

ACCEPTED FOR PUBLICATION
22 May 2025

PUBLISHED
11 June 2025

Abstract

Deep generative models are a powerful tool for exploring the chemical space within inverse-design workflows; however, their effectiveness relies on sufficient training data and effective mechanisms for guiding the model to optimize specific properties. We demonstrate that designing an expert-informed data representation and training procedure allows leveraging data augmentation while maintaining the required sampling controllability. We focus our discussion on a specific class of compounds (transition metal complexes), and a popular class of generative models (equivariant diffusion models), although we envision that the approach could be extended to other chemical spaces and model types. Through experiments, we demonstrate that augmenting the training database with generic but related unlabeled data enables a practical level of performance to be reached.

1. Introduction

Accelerating the discovery of novel compounds with desired properties is a grand challenge, but searching the chemical space is notably difficult—the two main challenges being: how to explore the space and how to evaluate properties? Experimental synthesis and testing are impractical, and the computational cost of *in-silico* screening with *ab-initio* methods is prohibitive at scale. While expensive calculations can be effectively amortized through surrogate models (Meyer *et al* 2018, Friederich *et al* 2020), the need for innovative exploration methods remains, and deep generative modeling has recently emerged as a promising tool for this task (Anstine and Isayev 2023). This type of model is capable of learning complex data distributions, that, in turn, can be sampled from to obtain novel compounds sharing similarities with the training database (Ruthotto and Haber 2021). While promising on paper, the potential success of such models hinges on (1) the availability of adequate training data in a sufficient amount, (2) the possibility of controlling the sampling process, as the search process often needs to focus on subsets of the full design space. Depending on the problem at hand, collecting large amounts of data can be challenging, and reliable controllability is difficult to achieve, preventing deep generative models to be satisfactorily trained and included in inverse-design workflows.

To mitigate the chronic issue of data scarcity in computational chemistry, we take inspiration from data augmentation. It is a common technique used to enhance the performance of machine learning models without requiring additional labeled data—supplemental training samples are obtained by altering existing ones through *augmentation functions*. While mostly employed in discriminative settings, augmentation functions can also be used in generative modeling when properly set up (Jun *et al* 2020). In an inverse-design context, the goal is to maximize success rate, a quantity that, regardless of the property of interest, directly depends on the validity and novelty of the generated samples—the model should generate novel chemically valid compounds featuring some desired property. In opposition to the usual data augmentation setting, augmentation functions that simply alter existing samples are unlikely to lead to increased novelty. Instead,

by exploiting generic (but related) datasets as augmentations, the model can learn and transfer knowledge from the augmented samples—e.g. transfer local chemical motifs. In the sequel, we focus our exposition on transition metal complexes (TMCs), a class of compounds especially relevant in homogeneous catalysis (Nandy *et al* 2021), where the discovery of novel high-performing catalysts for critical chemical reactions is of high importance. To determine the effectivity of a candidate TMC catalyst, expensive density functional theory (DFT) calculations (Kohn *et al* 1996), e.g. involving transition state (TS) searches, are frequently required. Task-specific datasets are therefore often limited in size. Moreover, due to the structure of TMCs, i.e. a TM center surrounded by ligands coordinated in specific ways, datasets are commonly constructed by combining a small number of ligands (Meyer *et al* 2018, Friederich *et al* 2020, Krieger *et al* 2021)—thereby increasing further the risk of overfitting. In the realm of TMCs, several practically relevant tasks require some form of controllability; for instance to enforce specific constraints on the generated TMCs: specific metal center, coordination pattern or ligand denticity. Or alternatively, to combine prior knowledge, e.g. via a known seed compound, along with the ability of generative models to suggest novel complexes by fixing the metal center and some of the ligands, and asking the model to generate the remaining ones.

In this work, we design an expert-informed data representation of TMCs, and formulate a conditional generative model. We resort to an equivariant diffusion model operating in 3D (Hoogeboom *et al* 2022, Jin and Merz 2024a, 2024b, Cornet *et al* 2024b), as it has been used in related tasks where the geometry is also of great importance (Guan *et al* 2022, Igashov *et al* 2024, Schneuing *et al* 2024). We note that similar models operating on geometric graphs (Daigavane *et al* 2024, Irwin *et al* 2024, Qu *et al* 2024, Cornet *et al* 2024a) could in principle fit equally well. The proposed data representation allows large generic databases, *e.g.* including different coordination patterns, to be leveraged for training while providing the required controllability. We employ the trained model and demonstrate its potential on two practical tasks: (1) generation from scratch followed by a screening campaign, and (2) exploration of the chemical space around a known complex via redesign of one of its constituting ligands.

2. Case study

2.1. Dataset

As a representative example for our case study, we select the dataset from Friederich *et al* (2020), which covers the chemical space around the well-studied Vaska's complex (Vaska and DiLuzio 1961). In this system, the coordinated ligands activate the Ir center for the oxidative addition of hydrogen. The corresponding reaction is graphically depicted in figure 1(a). The quantity of interest is the activation energy associated with the oxidative addition of hydrogen H_2-activation barrier: $\Delta E^{\ddagger}_{H_2}$, as it conditions the reaction rate. This quantity is expressed as a difference between the TS energy and the energy of the pair initial catalyst and H_2 group.

To define the chemical space around Vaska's complex, Friederich *et al* (2020) generalized the original formula, $[Ir(PPh_3)_2\,(CO)(Cl)]$, to the more generic $[IrA_2BC]$, where each group (i.e. A, B, C) got populated with a number of different ligands—all illustrated in figure 1(b). The ligands directly influence the activation barrier of the reaction due to the virtue of their $\sigma-/\pi-$electron donor or acceptor characters. The ligands belonging to group A are of $\sigma-$donor character, while ligands from group B can be either $\sigma-$ or $\pi-$donors. In group C, ligands can be $\sigma-$donors or $\pi-$acceptors. The database was constructed by enumerating all possible combinations of ligands in the *trans* Ir(I) square planar framework, yielding a total of 2574 unique complexes. Out of these, TS calculations were successfully performed for 1947 complexes. In appendix B, we perform a handful of calculations for structures from the original dataset (Friederich *et al* 2020), to ensure that our DFT setup is in agreement with the original one.

2.2. Task

In our experiments, we seek to directly generate guesses of TS geometries (i.e. structures on top of the energy profile in figure 1(a)). While Friederich *et al* (2020) successfully trained an accurate predictive model on it, the dataset is not ideal for training a deep generative model. Its limited size and diversity are likely to incur low validity and overfitting. It therefore constitutes a good test bench, where the target distribution is really application-specific—as we are unlikely to find similar data elsewhere, being specific to this reaction.

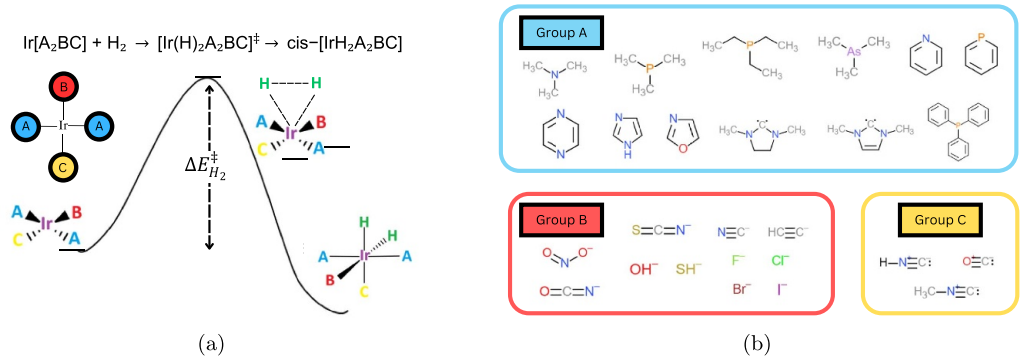


Figure 1. (a) Reaction under study, along with the generalized Vaska's complex [IrA_2BC], and the target property $\Delta E^{\ddagger}_{H_2}$ considered in this work. (b) Ligands (26 in total) used to construct the initial dataset (Friederich *et al* 2020), grouped by type.

3. Methods

3.1. Data augmentation via conditional generative modeling

In the supervised learning setting, where the goal is to learn the conditional distribution of a response variable y given an input variable $\boldsymbol{x}$: $p(y|\boldsymbol{x})$, data augmentation is a common technique to improve model performance without requiring additional (expensive) labeled data. The basic idea is to create (cheap) supplemental training examples from existing ones, using so-called *augmentation functions* (Zhang *et al* 2018, Cubuk *et al* 2019). The main hypothesis behind this procedure is that a careful design of augmentation functions can act as an inductive bias, and lead to better generalization of the overall model (Shorten and Khoshgoftaar 2019). Depending on the problem at hand, many augmentation functions, $a(\cdot)$, can potentially be designed, such that $p\big(y|a(\boldsymbol{x})\big)$ is left unchanged. In image classification, corrupting images with moderate amounts of noise or random rotations is common practice, as these transformations leave the target class unchanged. Although similar procedures can be relevant in computational chemistry too—as demonstrated by Godwin *et al* (2022) who improved model generalization on energy-related properties by extending the training data with perturbed geometries combined with an auxiliary denoising training objective, or Hu *et al* (2021) that used rotation augmentations to encourage rotation equivariance of an unconstrained neural network interatomic potential; constructing data augmentations is in general less straightforward and requires domain knowledge, e.g. to design meaningful substitutions of similar chemical elements or substructures (Magar *et al* 2022). Additionally, depending on the property of interest, augmentations can have undesirable consequences, as small perturbations in $\boldsymbol{x}$ can lead to significant changes in y (Stumpfe *et al* 2019). These activity cliffs must be taken into account when crafting augmentations, in order not to create augmented (and seemingly similar) samples with vastly different chemical properties (Van Tilborg *et al* 2022).

In generative modeling, the goal is different and instead consists in modeling a target distribution $p_{\text{target}}(\boldsymbol{x})$ (e.g. blue distribution in figure 2) given a finite amount of samples thereof, $\mathcal{D}_{\text{target}}$. When data augmentation is performed, we additionally have access to auxiliary samples, $\mathcal{D}_{\text{aug.}}$, coming from an augmentation distribution $p_{\text{aug.}}(\boldsymbol{x})$ (green distribution in figure 2), e.g. implicitly defined by transforming $\mathcal{D}_{\text{target}}$ through an augmentation function $a(\cdot)$. If the two data sources, $\mathcal{D}_{\text{target}}$ and $\mathcal{D}_{\text{aug.}}$, are simply mixed together as in the supervised setting, the model can not learn to distinguish between them and end up modeling an approximate joint distribution, such as the red distribution depicted in figure 2. At sampling time, the trained model is then likely to generate samples that are potentially very unlikely under the target distribution $p_{\text{target}}(\boldsymbol{x})$ – e.g. noisy images. Instead, the key idea is to enable the model to distinguish between the different distributions (Jun *et al* 2020). This can be done by modeling a conditional distribution: $p_\theta(\boldsymbol{x}|\boldsymbol{c})$ instead of $p_\theta(\boldsymbol{x})$, as illustrated in the right inset in figure 2—where $\boldsymbol{c}$ is a conditioning vector distinguishing between original and augmented samples. As a single model is trained to approximate both distributions through conditioning, it can benefit from the augmentation while remaining controllable at sampling time. In the simplified example from figure 2, we for instance expect the model to approximate the data distribution along the $x-$axis better after augmentation. Samples from the target distribution p_{target} can simply be obtained by conditioning on $\boldsymbol{c}_{\text{target}}$: $p_{\text{target}} \approx p_\theta(\cdot|\boldsymbol{c}_{\text{target}})$.

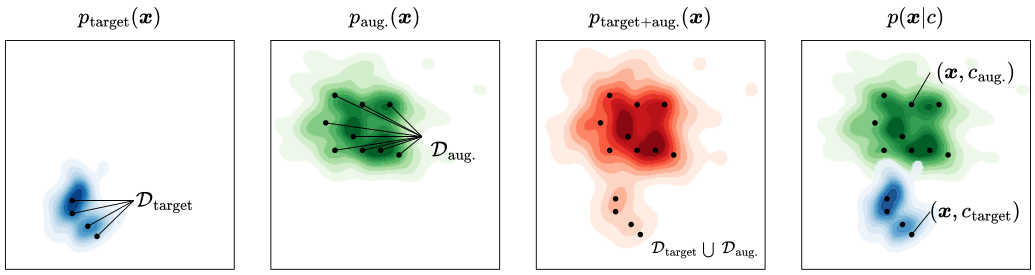


Figure 2. Cartoon representation of the data augmentation setting in generative modeling. By simply augmenting the training data with samples coming from an augmentation distribution, samples that are unlikely under the target distribution $p_{\text{target}}(\boldsymbol{x})$ are going to be generated. Instead, by modeling a conditional distribution, $p_\theta(\boldsymbol{x}|\boldsymbol{c})$, data augmentation can be leveraged while allowing for proper sampling controllability.

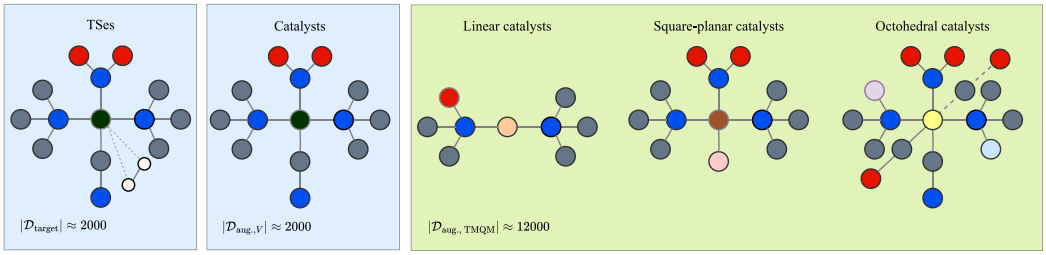


Figure 3. Schematic view of the data augmentation procedure. The initial set of TS structures is augmented with isolated catalysts, coming from the original (Friederich *et al* 2020) and a generic database of TMCs (Balcells and Skjelstad 2020), and featuring different coordination patterns.

3.2. Data augmentation for TMCs

As touched upon in section 3.1, data augmentation aims at improving generalization. Starting from a limited dataset as the one described in section 2, our goal is to devise a careful data augmentation scheme to instill additional chemical knowledge and improve the capabilities of the generative model. In this setting, standard augmentation functions altering existing samples are not applicable—e.g. creating additional samples by perturbing the geometry of existing structures is unlikely to equip the model with more chemical knowledge. Instead, we propose a simple approach consisting in augmenting the training data with samples coming from a generic, but related, dataset. In particular, we acquire supplemental training examples from the updated TMQM dataset (Balcells and Skjelstad 2020), a generic database of TMCs.

Concretely, we augment the original dataset (Friederich *et al* 2020) with selected complexes extracted from TMQM (Balcells and Skjelstad 2020). We filter the database, and keep (1) linear, square-planar and octohedral coordination patterns as inferred by `MolSimplify` (Ioannidis *et al* 2016, Nandy *et al* 2018) – i.e. most similar patterns to the target square-planar; (2) monodentate ligands only—as required by the problem at hand; (3) TMCs consisting of 100 atoms at most—to limit the GPU memory requirements. After filtering, around 12000 supplemental training samples are added to the training data. Finally, to enable the model to learn the relationship between a relaxed catalyst and its corresponding TS for the reaction presented in figure 1(a), we additionally add the relaxed catalysts from the original dataset (Friederich *et al* 2020). In total, the augmented training set now comprises 16 000 data points, of which around 2000 are TS structures. The augmentation procedure is graphically summarized in figure 3. We also provide a low-dimensional visualization of the effect of the augmentation procedure in appendix A and figure 8. We note that alternative data curation processes could be designed, and that other large-scale databases, e.g. CSD (Groom *et al* 2016), would be equally relevant.

3.3. Expert-informed data representation of TMCs

Formally, we represent a N-atom TMC as a tuple $(\boldsymbol{x}, \boldsymbol{c})$, where $\boldsymbol{x} = (\boldsymbol{r}, \boldsymbol{h})$ is the usual geometric graph representation (Hoogeboom *et al* 2022), with $\boldsymbol{r} = (\boldsymbol{r}_1, \ldots, \boldsymbol{r}_N) \in \mathbb{R}^{N\times 3}$ and $\boldsymbol{h} = (\boldsymbol{h}_1, \ldots, \boldsymbol{h}_N) \in \mathbb{R}^{N\times D}$ denoting atomic coordinates and features respectively. To effectively leverage the augmented dataset presented in section 3.2 using the ideas introduced in section 3.1, we need to design the conditional information vector $\boldsymbol{c}$ such that it enables (1) the handling of different coordination patterns and ligand denticities, (2) generation

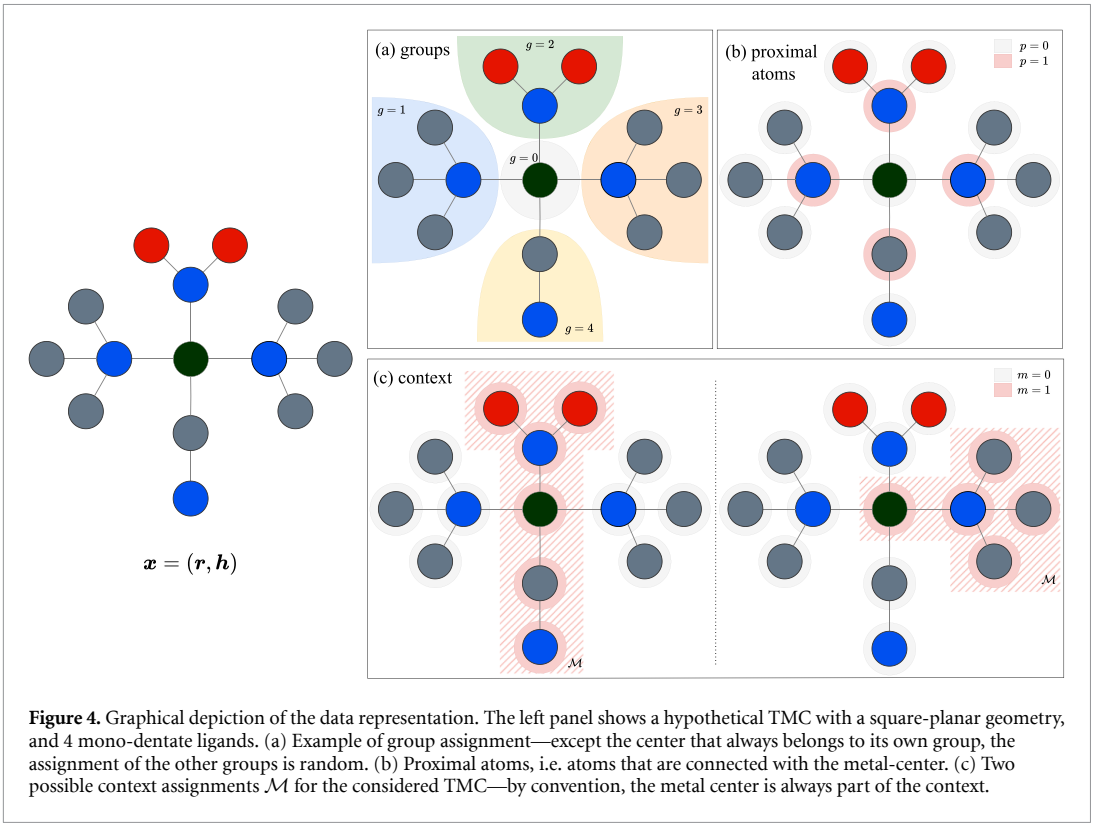


Figure 4. Graphical depiction of the data representation. The left panel shows a hypothetical TMC with a square-planar geometry, and 4 mono-dentate ligands. (a) Example of group assignment—except the center that always belongs to its own group, the assignment of the other groups is random. (b) Proximal atoms, i.e. atoms that are connected with the metal-center. (c) Two possible context assignments $\mathcal{M}$ for the considered TMC—by convention, the metal center is always part of the context.

of TS structures on demand—our target. The model should therefore use the conditional information to factor the different coordination patterns and ligand denticities out of its internal representation (i.e. this information now becomes an input), and to distinguish between TS and non-TS structures. Additionally, as we would like to perform context-conditioned generation, the conditional vector $\boldsymbol{c}$ should also allow the model to identify which atoms are considered as context. We therefore resort to atom-level information $\boldsymbol{c} = (\boldsymbol{c}_1, \ldots, \boldsymbol{c}_N) \in \mathbb{R}^{N \times C}$ defined as follows,

$$\boldsymbol{c} = [\boldsymbol{g} || \boldsymbol{p} || \boldsymbol{m}]\,,$$

with $||$ denoting concatenation, $\boldsymbol{g} \in \{0, 1, \ldots, G\}^N$ gathering the group information, $\boldsymbol{p} \in [0, 1]^N$ specifying which atoms are connected to the metal center, and $\boldsymbol{m} \in [0, 1]^N$ encoding which atoms are part of the context. A graphical depiction of the data representation and the different variables is presented in figure 4.

Group and connectivity information, i.e. $\boldsymbol{g}$ and $\boldsymbol{p}$, enable the model to discern the different coordination patterns and denticities. Specifically, $\boldsymbol{g}$ specifies which group, i.e. center or ligand, an atom belongs to—as illustrated in inset (a) in figure 4. By convention, the center always belongs to group 0. The variable $\boldsymbol{p}$ indicates which atoms are proximal, i.e. connected to the metal center—as depicted in inset (b) in figure 4. In the augmented dataset as we only have monodentate ligands, there can be either 3, 5 or 7 groups—respectively corresponding to linear, square-planar and octohedral coordination patterns, where there is exactly one atom in each group that is proximal.

Context-conditioned generation is enabled by $\boldsymbol{m}$, that partitions the atoms into two subsets: $\boldsymbol{x} = \{\boldsymbol{x}^{\in \mathcal{M}}, \boldsymbol{x}^{\notin \mathcal{M}}\}$, where subset $\mathcal{M}$ is treated as context. The metal center is always part of $\mathcal{M}$. An example of two possible context assignments for a given complex is illustrated in inset (c) in figure 4.

3.3.1 Representing TSes

The target is to generate TS guesses directly. We therefore need to prompt the model with the right conditioning $\boldsymbol{c}_{\text{target}}$ to obtain samples from that distribution. As in the problem under study, TSes include an additional H_2 group compared to isolated catalysts, $\boldsymbol{c}_{\text{target}}$ is designed by adding two atoms to group 0, i.e. with the Ir metal center, and making these two atoms proximal. As a concrete example, the TS corresponding to the TMC depicted in figure 4 would feature 5 groups: a center and four ligand groups. As all considered ligands are monodentate, i.e. exactly one atom binding to the metal center would be considered as proximal in each of the ligand groups. The center group, $\boldsymbol{g} = 0$, would additionally include two atoms representing H_2, in which both atoms would be considered proximal, $\boldsymbol{p} = 1$.

3.4. Generative model

3.4.1. Conditional equivariant diffusion model

To exploit the data augmentation procedure introduced in section 3.1 and perform context-conditioned generation, we formulate a conditional probability distribution $p_\theta(\boldsymbol{r}^{\notin\mathcal{M}}, \boldsymbol{h}^{\notin\mathcal{M}} | \boldsymbol{r}^{\in\mathcal{M}}, \boldsymbol{h}^{\in\mathcal{M}}, \boldsymbol{c})$. We employ an equivariant diffusion model similar to that of OM-DIFF (Cornet *et al* 2024b), where the forward and reverse processes only alter the atomic positions and features of non-contextual atoms $(\boldsymbol{r}^{\notin\mathcal{M}}, \boldsymbol{h}^{\notin\mathcal{M}})$, whereas $\boldsymbol{c}$ and the contextual atomic positions and features $(\boldsymbol{r}^{\in\mathcal{M}}, \boldsymbol{h}^{\in\mathcal{M}})$ are left unchanged. The reverse process is parameterized by a conditional equivariant denoising network $\boldsymbol{\varepsilon}_\theta(\boldsymbol{x}_t, t, \boldsymbol{c})$, with $\boldsymbol{x}_t = (\boldsymbol{r}_t, \boldsymbol{h}_t)$ denoting the noisy version of $\boldsymbol{x}$ at time t. The noisy atomic features $\boldsymbol{h}_t$ and the conditional information $\boldsymbol{c}$ get concatenated, and processed through a shared encoder to obtain the initial invariant hidden states for each atom in the complex.

3.4.2. Multi-task conditional training

For each data sample, $\boldsymbol{g}$ and $\boldsymbol{p}$ are fixed but $\boldsymbol{m}$, and thereby $\mathcal{M}$, are yet to be defined, depending on what part of the complex should be considered as context. We therefore construct $\boldsymbol{m}$ on-the-fly, by drawing a random combination of ligand groups to be masked. A given complex $\boldsymbol{x}$ appears thus multiple times during training, with different variations of $\boldsymbol{c}$. That way, $\boldsymbol{\varepsilon}_\theta$ is trained to denoise structures with varying contexts. This can be seen as a form of multi-task training, where the model is trained to approximate several distributions. We note that, depending on the downstream task, $\boldsymbol{m}$ could also be constructed differently, e.g. at atom-level instead of ligand-level.

3.4.3. Conditional sampling of TSes

During training, the model sees several coordination patterns, a mixture of isolated catalysts and TS structures, and varying contexts. At sampling time, the model is provided with the desired conditional information $\boldsymbol{c}_{\text{target}}$, as to generate TSes. Namely, we specify the Ir metal center, two atoms belonging to the same group as the center representing H_2, where each H is proximal. Finally, we create 4 groups of monodentate ligands. To perform context-conditioned generation around a known complex, $(\boldsymbol{r}^{\in\mathcal{M}}, \boldsymbol{h}^{\in\mathcal{M}})$ and $\boldsymbol{m}$ can be specified accordingly, but the H_2 group is never part of the context.

4. Experiments and results

In what follows, we perform two experiments to demonstrate the benefits afforded by the data representation, the training and augmentation procedures. In section 4.1., we sample novel complexes from scratch, yielding a broad coverage of the property space. In addition to evaluating the model, this can be practically useful to *e.g.* extend an existing database or to perform screening. In section 4.2, we leverage the ability of the model to perform context-conditioned generation and concentrate the sampling around a known promising complex. We demonstrate that it effectively allows the property space to be searched locally.

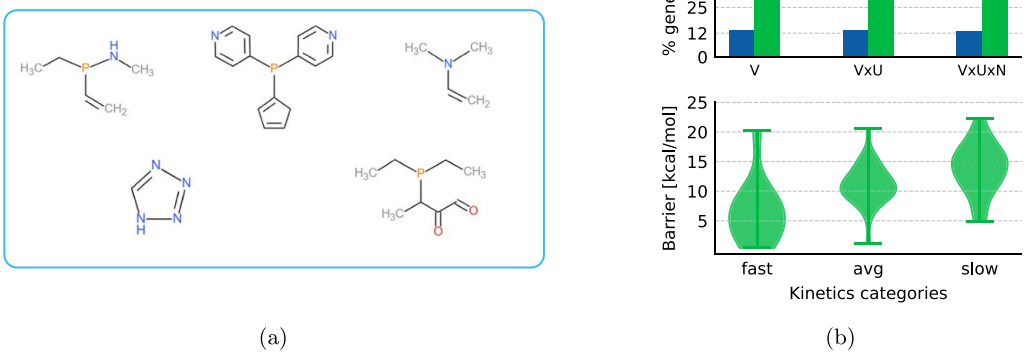
4.1. Sampling from scratch

4.1.1. Setup

We construct the necessary conditional information to generate a TS $\boldsymbol{c}_{\text{target}}$ as described in section 3.3.1, and sample 10 000 complexes. The number of atoms composing the different ligands is drawn as follows: $N_A \sim \mathcal{U}([5,40])$ $N_B \sim \mathcal{U}([1,6])$ and $N_C \sim \mathcal{U}([2,6])$—where A, B and C refer to the ligand groups illustrated in figure 1(a). We evaluate the samples generated in terms of validity, uniqueness and novelty. In short, a generated sample is deemed valid if the H_2 distance and the distance Ir–H_2 are reasonable, and if each of the ligands can be sanitized by `rdkit` (Landrum 2024). Uniqueness and validity are evaluated by converting the ligands to SMILES, and treating complexes as multisets (Cornet *et al* 2024b) when comparing them. We provide additional details about the evaluation procedure in appendix D.1.

4.1.2. Impact of data augmentation

First, we seek to quantify the impact of data augmentation. To do so, we compare two variants of the same model: one variant trained on TS structures only, as provided in the target dataset (Friederich *et al* 2020), and another variant trained on the augmented dataset detailed in section 3.2. In the top panel of figure 5(b), we observe that augmenting the training data with non-TS geometries, including other coordination patterns leads to a clear improvement of the generative capabilities of the model—an approximately 4-fold improvement, increasing the the proportion of $V \times U \times N$ complexes from around 12% to nearly 50%. We note that for each model variant, nearly all valid samples are both unique and novel.



(a) (b)

Figure 5. (a) Excerpt of 5 novel A ligands obtained in the generation from scratch experiment. (b—top) Augmenting the dataset greatly increases Validity, Uniqueness and Novelty, approximately a 4-fold improvement. (b—bottom) DFT energy barrier distribution for the kinetics categories estimated by the surrogate model.

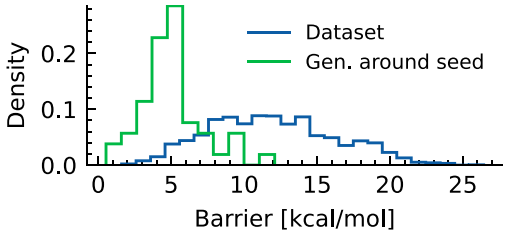


Figure 6. Results of the local search experiment. Without conditioning directly on the barrier energy, searching in the neighborhood of a seed complex, with a barrier energy of 1.6 kcal mol^{-1}, leads to a shift in the property distribution, as per DFT.

4.1.3. Data-driven screening

We train a surrogate model (more details in appendix C), and use it to estimate the barrier energy for the $V \times U \times N$ samples generated by the model trained on the augmented dataset. The model predictions are used to categorize each sample according to its kinetics as per Friederich *et al* (2020): fast ($\Delta E^{\ddagger}_{H_2} < 8.1$ kcal mol^{-1}), average (8.1 kcal mol^{-1} $< \Delta E^{\ddagger}_{H_2} <$ 15.1 kcal mol^{-1}), or slow ($\Delta E^{\ddagger}_{H_2} > 15.1$ kcal mol^{-1}). We then run DFT calculations (more details in appendix B) to compute the true barrier of $\approx$ 400 samples, of which 120 were successful. We discuss the failure modes in more details in appendix D.2. For each kinetics category, we display the corresponding energy distributions in the bottom panel of figure 5. While not perfect, we observe that the surrogate is capable of capturing the overall energy trend of the complexes generated by the diffusion model, indicating that a similar fully data-driven procedure could potentially be used in a real discovery pipeline—e.g. to quickly discard or select complexes that belong to a kinetics category or another.

4.2. Searching in the neighborhood of a known complex

4.2.1. Setup

We again provide $\boldsymbol{c}_{\text{target}}$ to generate a TS structure, but we now also input a context set $\boldsymbol{x}^{\mathcal{M}}$ corresponding to the complex with lowest activation energy in the dataset (Friederich *et al* 2020), where one of the A ligands has been removed—we provide a graphical illustration of the setting in the upper left corner in figure 7. The model is tasked to redesign the removed A ligand (including the coordinating atom), and place H_2 accordingly. The size of the ligand to be designed is drawn from $N_A \sim \mathcal{U}([5, 40])$.

4.2.2. Local search

We collect and run DFT calculations for around 200 randomly selected $V \times U \times N$ samples. Among these, 50 calculations were successful (more details in appendix D.2). The resulting barrier energy distribution is provided in figure 6, where we can observe a clear shift towards lower barrier energies, compared to the initial dataset distribution. This shift can be explained by the low barrier energy (1.6 kcal mol^{-1}) of the seed complex. We note that this shift is obtained without explicit conditioning on the property of interest, and is

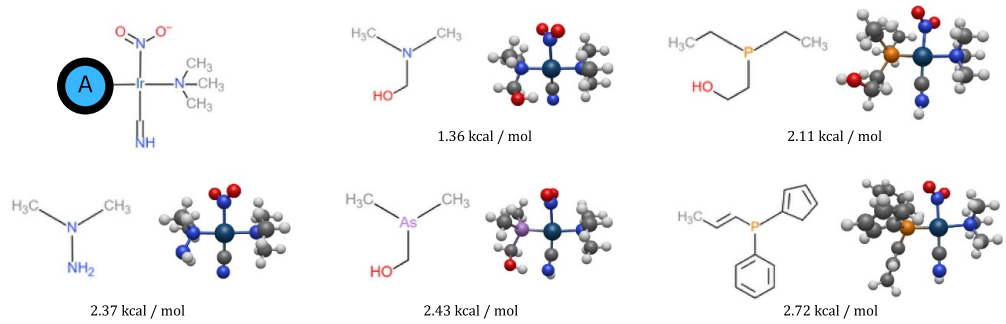


Figure 7. Novel complexes obtained by redesigning one of the A ligands in the seed complex extracted from the Vaska's dataset (Friederich *et al* 2020). The context $\boldsymbol{x}^{\mathcal{M}}$ provided to the model is schematically depicted in the upper left corner. Out of the 50 successful calculations, these 5 complexes featured the lowest barrier energies. Each complex is presented with the designed A ligand.

simply the result of alternative ligand designs suggested by the model, given the fixed context. Including the desired barrier energy into the conditioning scheme could potentially lead to an even sharper property distribution around the target value.

In figure 7, we provide a graphical depiction of the 5 novel samples that led to the lowest calculated energy barriers. We note that one of the complex has an activation barrier slightly lower than the lowest one found in the dataset. This experiment illustrates that local search around known promising complexes allows the combination of prior knowledge, i.e. through a known compound, along with the ability of generative models to suggest novel complexes.

5. Discussion and conclusion

In this paper, we targeted the chronic issue of data scarcity in generative modeling applied to computational chemistry, with a special focus on TMCs. To impart more knowledge to our generative model, we devised a tailored data representation allowing a series of conditional generation tasks relevant for TMC catalysts to be performed. In a case study revolving around Vaska's complex, we showed that the data representation enabled the training database to be augmented with generic data, thereby leading to a significant improvement in the generative capabilities of the model, while maintaining full controllability at test time. We then showed that the generative model could effectively be combined with a surrogate model to perform screening in a fully data-driven fashion. Finally, we demonstrated that the ability of the model to perform context-conditioned generation can be used to redesign parts of known complexes with desirable properties, and that the procedure is a viable approach to search the chemical space locally while inducing a desired shift in the property distribution.

While relatively modest in scope, we believe that our paper can provide a blueprint for enhancing generative model capabilities using domain knowledge and data from related tasks, beyond TMCs design. The proposed technique is simple and can easily be employed in other settings, with other model architectures and in other chemical spaces, where similar, albeit bespoke, representations could be designed. Additionally, our approach is orthogonal to existing methods that seek to improve capabilities of generative models through architectural and framework improvements, and can therefore be readily combined with these, as it focuses on the data and its representation—similarly to methods that additionally model bond or charge information (Vignac *et al* 2023).

Data availability statement

All data that support the findings of this study are included within the article (and any supplementary files).

Code availability statement

The code that support the findings of this study are openly available at the following URL/DOI: https://github.com/frcnt/om-diff (Cornet *et al* 2024b).

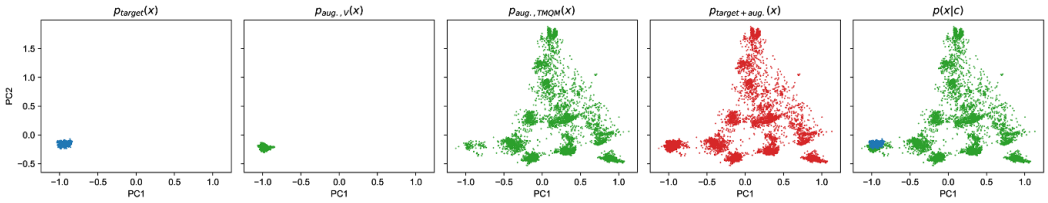


Figure 8. Visualization of the dataset before and after data augmentation. The fixed-size descriptors are obtained by (1) embedding all configurations using `MACE-MP-medium` (Batatia *et al* 2023), (2) keeping the final metal-center descriptors for each configuration, (3) projecting the vectors using PCA. The two first principal components explain 76% of the variance.

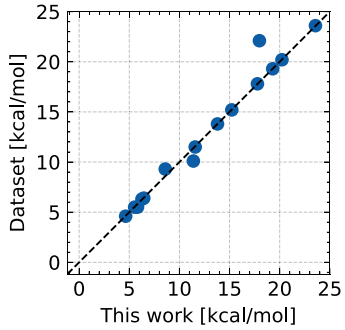


Figure 9. Parity plot showing comparing the energies obtained using our DFT setup against the values reported in the original dataset (Friederich *et al* 2020).

Appendix A. Effect of data augmentation

The hypothesis behind data augmentation is that the generalization of the model along certain dimensions can be improved by learning from augmented samples. For instance, in the cartoon representation in figure 2, we can expect the model to benefit from the augmentation along the $x-$axis primarily. In high dimensions, the picture is a bit more nuanced, but we expect the model to gain additional general chemical knowledge via data augmentation, as confirmed by the improved validity and novelty metrics in figure 5(b).

In figure 8, we attempted to visualize the effect of data augmentation on our dataset. To represent the data in a low-dimensional space, we (1) obtained embeddings for all configurations using `MACE-MP-medium` (Batatia *et al* 2023), (2) kept the final metal-center descriptors for each configuration, (3) performed PCA and projected the descriptors along the two first principal components. We observe that the augmentation provides a variety of local environments that were not present in the original dataset, while showing a pattern along `PC2` similar to that of the cartoon representation along x. We however note that this provides an oversimplified view, as the dataset is solely described from the perspective of the metal center.

Appendix B. DFT setup

A DFT protocol is set up to obtain the relaxed geometry of the catalyst, search for the TS and get converged energies of both the structures which are then used to calculate the activation barrier. For the geometry optimization of a catalyst, we use the TS generated by the model, remove the activated hydrogens from the metal center and use the resulting structure as the initial guess. The level of theory used is the same as the one used in the original dataset. The calculations were performed in Gaussian16 (Frisch *et al* 2016) with the PBE functional (Perdew *et al* 1996). Optimizations were performed with the def2-SVP basis set (Schäfer *et al* 1992). Effective core potentials are specified for the non-valence electrons of iridium. Furthermore, Grimme's D3 (Grimme *et al* 2010) dispersion correction was used. Optimizations were performed with a convergence criterion `tight` in Gaussian. The TS search was carried out at the same level of theory (figure 9). First the H–H bond was frozen at 1 Å and rest of the molecule was optimized, following this the optimized structure was used as a starting guess for TS search. Frequency calculations were carried out to confirm that the obtained TS corresponded to H–H bond breaking.

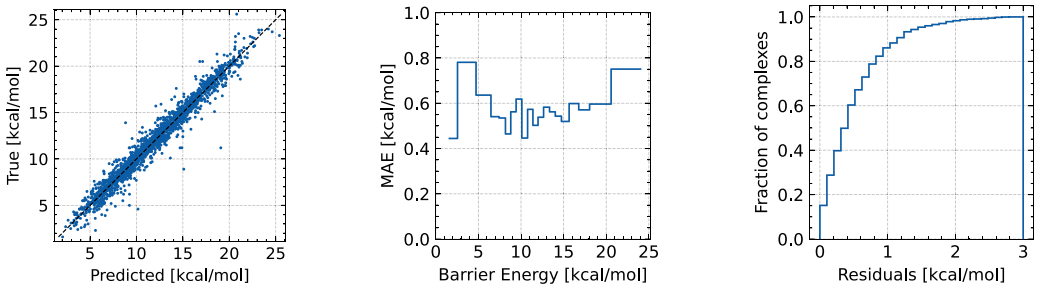


Figure 10. Error diagnostic of the surrogate model. (a) Parity plot. (b) MAE across the property space. (c) Cumulative distribution of the residuals.

Table 1. Numerical summary of the error diagnostic of the surrogate model obtained by performing 10-fold cross-validation. Errors are provided in kcal · mol^{-1}. Standard deviation is computed across folds.

Metric	Value
MAE (↓)	$0.58_{\pm 0.04}$
RMSE (↓)	$0.83_{\pm 0.09}$
MaxAE (↓)	$4.41_{\pm 1.62}$
R^2 (↑)	$0.96_{\pm 0.01}$

Appendix C. Surrogate model

As it is not practical to use DFT to compute energy barriers for all the samples produced by the generative model, we resort to a surrogate model to perform a cheap screening. We use another equivariant neural network model, and train it TS structures from Friederich *et al* (2020). We perform a 10-fold cross-validation to get an error estimate across the whole property space. The corresponding diagnostic can be seen in figure 10 and table 1.

Appendix D. Evaluation details

D.1. Checks

D.1.1. Validity check

To evaluate the validity of the generated TS structures, we perform series of checks, that primarily rely on the ability of `RDKit` (Landrum 2024) to infer bonding information:

1. **[pairwise H_2 distance check]** The pairwise distance the two H atoms, d_{H_2}, should be such that $d_{H_2} \in [0.7, 1.2]$ Å;
2. **[pairwise H_2-Ir distance check]** The pairwise distance the two H atoms and the Ir center, d_{H_2-Ir}, should be such that $d_{H_2-Ir} \in [1.5, 2.5]$ Å;
3. **[`RDKit` check]** We start by removing Ir and H_2 from the generated complex and we manually build a `Mol` object using the remaining atom types and coordinates,. We employ `rdDetermineBonds()` to infer the bond structure. As the overall charge of a catalyst should be 0, and that Ir has 1 positive charge, we allow bond assignments ligands that lead to a charge of −1. In principle, the negative charge should be on the B ligand, as shown in figure 1(b). After bond inference, the resulting `Mol` should contains 4 fragments, and `DetectChemistryProblems()` should return an empty list.

D.1.2. Uniqueness and novelty check

After a successful bond allocation for a complex, we convert each of its constituting fragments to a `SMILES` string, and finally represent the generated complex as a multi-set of strings. Uniqueness can be defined as the ratio of unique multi-sets among all generated samples, while novelty is expressed as the ratio of unique multi-sets that were not part of the training database.

D.2. Failure modes

While we only performed calculations on samples that were deemed valid (and novel) by the filters introduced appendix D.1, the TS guesses generated by the model are approximate (e.g. distorted structures

or stretched bonds), and can result in subsequent unsuccessful TS searches. Across our experiments, we observed a success rate of about 25% on average, against the ≈75% rate reported in Friederich *et al* (2020). We note that, in the latter, the TS were constructed by combining DFT-optimized ligand geometries.

The most common failure modes were the following:

- **Calculation timeout:** The calculations had an upper limit for run time of 10 h. Calculations that reached the time limit were instantly stopped and considered unsuccessful. About 50% of all the calculations did not complete within the specified time window;
- **Convergence failure:** This error indicates that the self-consistent field (SCF) procedure has failed to meet the convergence criterion. Solution to this error are dependent on the case such as changing the initial geometry, using a different level of theory or using a different SCF converger. This made up for 40% of the failed runs;
- **Torsional failure (`tors fail`):** This error is often encountered when using internal coordinates. This occurs when atoms line up in a straight line during the optimization process, since we are working with square planar geometries there is a often a chance of this situation arising and causing a failure. Torsonial failure accounted for 10% of the failed calculations.

ORCID iDs

François Cornet https://orcid.org/0009-0008-6157-862X
Pratham Deshmukh https://orcid.org/0009-0004-1719-7310
Bardi Benediktsson https://orcid.org/0000-0002-1578-9126
Mikkel N Schmidt https://orcid.org/0000-0001-6927-8869
Arghya Bhowmik https://orcid.org/0000-0003-3198-5116

References

Anstine D M and Isayev O 2023 Generative models as an emerging paradigm in the chemical sciences *J. Am. Chem. Soc.* **145** 8736–50
Balcells D and Skjelstad B B 2020 tmQM dataset—quantum geometries and properties of 86k transition metal complexes *J. Chem. Inf. Model.* **60** 6135–46
Batatia I *et al* 2023 A foundation model for atomistic materials chemistry (arXiv:2401.00096)
Cornet F R J, Bartosh G, Schmidt M N and Naesseth C A 2024a Equivariant neural diffusion for molecule generation *Advances in Neural Information Processing Systems* p 37 (available at: https://proceedings.neurips.cc/paper_files/paper/2024/hash/587b3f360588143a751c37fcb3b5db7f-Abstract-Conference.html)
Cornet F, Benediktsson B, Hastrup B, Schmidt M N and Bhowmik A 2024b Om-diff: inverse-design of organometallic catalysts with guided equivariant denoising diffusion *Digit. Discov.* **3** 1793–811
Cubuk E D, Zoph B, Mané D, Vasudevan V and Le Q V 2019 Autoaugment: Learning augmentation strategies from data *IEEE Conf. on Computer Vision and Pattern Recognition, CVPR 2019, (Long Beach, CA, USA, 16-20, 2019, 113–123)* (Computer Vision Foundation / IEEE) https://doi.org/10.1109/CVPR.2019.00020 (available http://openaccess.thecvf.com/content_CVPR_2019/html/Cubuk_AutoAugment_Learning_Augmentation_Strategies_From_Data_CVPR_2019_paper.html)
Daigavane A, Kim S E, Geiger M and Smidt T 2024 Symphony: symmetry-equivariant point-centered spherical harmonics for 3D molecule generation *The 12th Int. Conf. on Learning Representations* (available at: https://openreview.net/forum?id = MIEnYtlGyv)
Friederich P, dos Passos Gomes G, De Bin R, Aspuru-Guzik A and Balcells D 2020 Machine learning dihydrogen activation in the chemical space surrounding vaska's complex *Chem. Sci.* **11** 4584–601
Frisch M J *et al* 2016 *Gaussian16 Revision C.01* (Gaussian Inc)
Godwin J, Schaarschmidt M, Gaunt A L, Sanchez-Gonzalez A, Rubanova Y, Veličković P, Kirkpatrick J and Battaglia P 2022 Simple GNN regularisation for 3D molecular property prediction and beyond *Int. Conf. on Learning Representations* (available at: https://openreview.net/forum?id = 1wVvweK3oIb)
Grimme S, Antony J, Ehrlich S and Krieg H 2010 A consistent and accurate *Ab Initio* parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu *J. Chem. Phys.* **132** 154104
Groom C R, Bruno I J, Lightfoot M P and Ward S C 2016 The cambridge structural database *Struct. Sci.* **72** 171–9
Guan J, Qian W W, Peng X, Su Y, Peng J and Ma J 2022 3D equivariant diffusion for target-aware molecule generation and affinity prediction *11th Int. Conf. on Learning Representations* (available at: https://openreview.net/forum?id=kJqXEPXMsE0)
Hoogeboom E, Satorras V G, Vignac C and Welling M 2022 Equivariant diffusion for molecule generation in 3D *Int. Conf. on Machine Learning* (PMLR) pp 8867–87 (available at: https://proceedings.mlr.press/v162/hoogeboom22a.html)
Hu W, Shuaibi M, Das A, Goyal S, Sriram A, Leskovec, Parikh D and Zitnick C L 2021 Forcenet: a graph neural network for large-scale quantum calculations (arXiv:2103.01436)
Igashov I, Stärk H, Vignac C, Schneuing A, Satorras V G, Frossard P, Welling M, Bronstein M and Correia B 2024 Equivariant 3D-conditional diffusion model for molecular linker design *Nat. Mach. Intell.* **6** 417–27
Ioannidis E I, Gani T Z H and Kulik H J 2016 molSimplify: a toolkit for automating discovery in inorganic chemistry *J. Comput. Chem.* **37** 2106–17
Irwin R, Tibo A, Janet J P and Olsson S 2024 Efficient 3D molecular generation with flow matching and scale optimal transport *ICML 2024 AI for Science Workshop* (available at: https://openreview.net/forum?id = CxAjGjdkqu)
Jin H and Merz K M Jr 2024a LigandDiff: de novo ligand design for 3D transition metal complexes with diffusion models *J. Chem. Theory Comput.* **20** 4377–84
Jin H and Merz K M Jr 2024b Partial to total generation of 3D transition-metal complexes *J. Chem. Theory Comput.* **20** 8367—77

Jun H, Child R, Chen M, Schulman J, Ramesh A, Radford A and Sutskever I 2020 Distribution augmentation for generative modeling *Proc. of the 37th Int. Conf. on Machine Learning* (*Proc. of Machine Learning Research* vol 119) ed H Daumé III and A Singh (PMLR) (available at: https://proceedings.mlr.press/v119/jun20a.html) pp 5006–19

Kohn W, Becke A D and Parr R G 1996 Density functional theory of electronic structure *J. Phys. Chem.* **100** 12974–80

Krieger A M, Sinha V, Kalikadien A V and Pidko E A 2021 Metal-ligand cooperative activation of hx (x = h, br, or) bond on mn based pincer complexes *Z. Anorg. Allg. Chem.* **647** 1486–94

Landrum G 2024 RDKit: Open-source cheminformatics (available at: www.rdkit.org/)

Magar R, Wang Y, Lorsung C, Liang C, Ramasubramanian H, Li P and Farimani A B 2022 Auglichem: data augmentation library of chemical structures for machine learning *Mach. Learn.: Sci. Technol.* **3** 045015

Meyer B, Sawatlon B, Heinen S, Von Lilienfeld O A and Corminboeuf C 2018 Machine learning meets volcano plots: computational discovery of cross-coupling catalysts *Chem. Sci.* **9** 7069–77

Nandy A, Duan C, Janet J P, Gugler S and Kulik H J 2018 Strategies and software for machine learning accelerated discovery in transition metal chemistry *Ind. Eng. Chem. Res.* **57** 13973–86

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

Perdew J P, Burke K and Ernzerhof M 1996 Generalized gradient approximation made simple *Phys. Rev. Lett.* **77** 3865–8

Qu Y, Qiu K, Song Y, Gong J, Han J, Zheng M, Zhou H and Ma W-Y 2024 Molcraft: structure-based drug design in continuous parameter space *41st Int. Conf. on Machine Learning* (https://doi.org/10.5555/3692070.3693767)

Ruthotto L and Haber E 2021 An introduction to deep generative modeling *GAMM-Mitteilungen* **44** e202100008

Schäfer A, Horn H and Ahlrichs R 1992 Fully optimized contracted Gaussian basis sets for atoms Li to Kr *J. Chem. Phys.* **97** 2571–7

Schneuing A *et al* 2024 Structure-based drug design with equivariant diffusion models *Nat. Comput. Sci.* **4** 899–909

Shorten C and Khoshgoftaar T M 2019 A survey on image data augmentation for deep learning *J. Big Data* **6** 1–48

Stumpfe D, Hu H and Bajorath J 2019 Evolving concept of activity cliffs *ACS Omega* **4** 14360–8

Van Tilborg D, Alenicheva A and Grisoni F 2022 Exposing the limitations of molecular machine learning with activity cliffs *J. Chem. Inf. Model.* **62** 5938–51

Vaska L and DiLuzio J W 1961 Carbonyl and hydrido-carbonyl complexes of iridium by reaction with alcohols. Hydrido complexes by reaction with acid *J. Am. Chem. Soc.* **83** 2784–5

Vignac C, Osman N, Toni L and Frossard P 2023 Midi: mixed graph and 3D denoising diffusion for molecule generation *Joint European Conf. on Machine Learning and Knowledge Discovery in Databases* (Springer) pp 560–76 (arXiv:2302.09048)

Zhang H, Cisse M, Dauphin Y N and Lopez-Paz D 2018 Mixup: beyond empirical risk minimization *Int. Conf. on Learning Representations* (available at: https://openreview.net/forum?id = r1Ddp1-Rb)