

# Equivariant conditional diffusion model for exploring the chemical space around Vaska's complex

**Cornet, François; Deshmukh, Pratham; Benediktsson, Bardi; Schmidt, Mikkel N.; Bhowmik, Arghya**

*Published in:*
Proceedings of the 38th Conference on Neural Information Processing Systems (NeurIPS 2024) - AI4Mat Workshop

*Publication date:*
2025

*Document Version*
Peer reviewed version

# Equivariant conditional diffusion model for exploring the chemical space around Vaska's complex

**François Cornet**[†*] **Pratham Deshmukh**[*]
**Bardi Benediktsson** **Mikkel N. Schmidt** **Arghya Bhowmik**
Technical University of Denmark
[†]frjc@dtu.dk

## Abstract

Generative modelling has recently emerged as a promising tool to efficiently explore the vast chemical space. In homogeneous catalysis, Transition Metal Complexes (TMCs) are ubiquitous, and finding better TMC catalysts is critical to a number of technologically relevant reactions. Evaluating reaction rates requires expensive transition state (TS) structure search, making traditional library-based screening difficult. Inverse-design of TMCs with a model capable of generating good TS guesses can lead to breakthroughs in catalytic science. We present such generative model herein. The model is an instance of an equivariant conditional diffusion model, and the key innovation lies in its specific data representation and training procedure, that allow generic databases (*e.g.* non-TS structures) to be leveraged at training time, while offering the desired controllability at sampling time (*e.g.* ability to generate TSs on demand). We demonstrate that augmenting the training database with generic (but related) data enables a practical level of performance to be reached. In a case study, our model successfully explores the chemical space around Vaska's complex, where the property of interest is the $H_2$-activation barrier, in two distinct settings: generation from scratch, and redesign of a specific ligand in a known TMC. In both cases, we validate a selection of novel samples with Density Functional Theory (DFT) calculations.

## 1 Introduction

Accelerated discovery of novel compounds with desired properties is a grand challenge. The social and economic impacts of the discovery of novel high-performing catalysts for critical chemical reactions are potentially immense. However, searching the chemical space is notably difficult. Experimental synthesis and testing are impractical, and the computational cost of *in-silico* screening with *ab-initio* methods is prohibitive for large set of molecules. While some of these computationally intensive calculations can effectively be amortized through surrogate models (Friederich et al., 2020), the need for innovative search methods remains, and deep generative models have recently emerged as a promising avenue (Anstine and Isayev, 2023). This type of models is capable of learning complex data distributions, that, in turn, can be sampled from to obtain novel compounds sharing similarities with the training database. Two main design choices are required when developing such models: the generative modelling paradigm and the data representation. For applications where the 3D geometry is particularly relevant, *e.g.* protein pocket-conditioned generation (Igashov et al., 2024), or expensive to obtain, *e.g.* ts generation (Duan et al., 2023), variants of Diffusion Models (DMs) (Ho et al., 2020) operating on geometric graphs (Hoogeboom et al., 2022) constitute a sensible option.

In this paper, we target Transition Metal Complexes (TMCs), a class of chemical compounds especially relevant in homogeneous catalysis (Nandy et al., 2021). As a case study, we focus on the

[*]These authors contributed equally to this work.

38th Conference on Neural Information Processing Systems (NeurIPS 2024) – AI4Mat Workshop.

chemical space in the neighbourhood of Vaska's complex (Friederich et al., 2020), with a target of optimizing the $H_2$-activation barrier: $\Delta E^{\ddagger}_{H_2}$. As this quantity directly depends on the geometry of the TS, we resort to an E(3)-equivariant diffusion model operating in 3D (Hoogeboom et al., 2022). We design a specific data representation, and formulate a conditional generative model that can be controlled to generate TS geometries on-demand at sampling time. The designed data representation also allows large generic databases, *e.g.* including non-TS structures or different coordination patterns, to be leveraged for training. This is particularly useful when task-specific datasets are limited in size, scarce, and most importantly if constructed from a small number of chemical motifs – likely to result in potential overfitting when training deep generative models.

We summarize our main contributions as follows:

1. We design a data representation that enables (i) the handling of different coordination patterns, (ii) structure-conditioned generation, and (iii) generation of TS structures on demand. We demonstrate the practical benefit of the representation by training an equivariant conditional diffusion model on an augmented dataset (TS and non-TS structures), and show the performance improvement compared to a dataset limited to TS structures only.
2. We leverage the trained model and demonstrate its potential on two practical tasks: (i) generation from scratch followed by a screening campaign, and (ii) exploration of the chemical space around a known complex via redesign of one of its constituting ligands.

**Related Work** Equivariant conditional models have been applied to a range of practically relevant problems such as conformer generation (Xu et al., 2021), linker design (Igashov et al., 2024), structure-based design (Schneuing et al., 2022), or target-aware design (Guan et al., 2022). Recent work (Jin and Merz Jr, 2024a;b; Cornet et al., 2024a) has also resorted to similar models for TMCs.

## 2 Methods

In Section 2.1, we detail the dataset, the generative task at hand along with the data augmentation procedure. In Section 2.2, we describe the proposed data representation, the conditional generative model along with the training and sampling procedures.

### 2.1 Dataset and Task

**Target chemical space** As a representative example for our case study, we select the dataset from Friederich et al. (2020), which studied the chemical space around Vaska's complex. In this system, the coordinated ligands activate the Ir center for the oxidative addition of hydrogen.Friederich et al. (2020) generalized the original formula, $[Ir(PPh_3)_2(CO)(Cl)]$, to the more generic $[IrA_2BC]$, and populated each group (i.e. A, B, C) with a number of different ligands. 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 TMCs. Out of these, transition state calculations were successfully performed for 1947 complexes.

The reaction under study, along with the structure of the considered catalysts, are depicted in Fig. 1. The ligands used by Friederich et al. (2020) are all illustrated in Fig. 5.

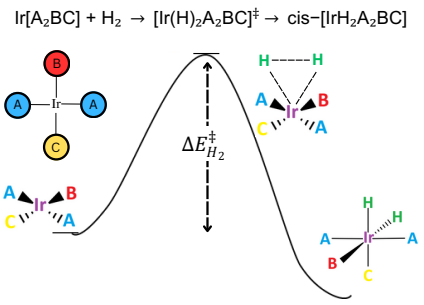


Figure 1: Reaction under study, along with the generalized Vaska's complex $[IrA_2BC]$, and the target property $\Delta E^{\ddagger}_{H_2}$ considered in this work.

**Task** The ultimate goal is to find complexes that minimize the activation energy associated with the oxidative addition of hydrogen $\Delta E^{\ddagger}_{H_2}$, as this is what conditions the reaction rate. With our model, we seek to directly generate good guesses of TS geometries (i.e. structures on top of the energy profile), in order to limit the amount of expensive transition state calculations required.

**Data augmentation** While enumerating a limited set of large *building blocks* is amenable to *ab-initio* screening, or a valid approach to building an accurate surrogate model (Friederich et al., 2020),

the resulting dataset is unlikely to be ideal for training a deep generative model, and the limited diversity is prone to incur overfitting. Ideally, such model should learn to generate novel structures while obeying strict chemical rules. Starting from a limited dataset, a careful data augmentation scheme can be designed to instill relevant chemical knowledge and improve the generative capabilities of the model.

We augment the original dataset (Friederich et al., 2020) with complexes extracted from TMQM (Balcells and Skjelstad, 2020), a database containing approximatively 108k TMCs. We filter the database, and keep (i) linear, square-planar and octohedral coordination patterns – the most similar patterns to the target square-planar; (ii) monodentate ligands only – as required by the problem at hand; (iii) TMCs consisting of 100 atoms at most – to limit the memory requirements. After filtering, around 12000 additional training samples are added to the training data. We provide additional details about the augmented dataset in Appendix A.2.

### 2.2 Generative Model

**Data representation** To effectively leverage augmented datasets – that include other coordination patterns, denticities and non-TS structures, while maintaining complete controllability at sampling time, the data representation should allow the model to (1) factor the different coordination patterns and ligand denticities out of its internal representation (i.e. this information should become an input), (2) distinguish between TS and non-TS structures, and (3) identify which atoms are considered as contextual information. With this in mind, we extend the usual geometric graph description with additional conditional information.

Formally, we represent a $N$-atom complex by $\boldsymbol{x} = (\boldsymbol{r}, \boldsymbol{h}, \boldsymbol{c})$, with $\boldsymbol{r} = (\boldsymbol{r}_1, ..., \boldsymbol{r}_N) \in \mathbb{R}^{N\times 3}$ and $\boldsymbol{h} = (\boldsymbol{h}_1, ..., \boldsymbol{h}_N) \in \mathbb{R}^{N\times D}$ denoting the usual atomic coordinates and features (*e.g.* atomic types or charges), along with $\boldsymbol{c} = (\boldsymbol{c}_1, ..., \boldsymbol{c}_N) \in \mathbb{R}^{N\times C}$ that specifies the conditional atomic information. Importantly, $\boldsymbol{c}$ is defined as follows,

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

where $||$ denotes concatenation, $\boldsymbol{g} \in \{1, ..., G\}^N$ gathers the ligand group information (i.e. which ligand an atom belongs to), $\boldsymbol{p} \in [0, 1]^N$ specifies which atoms are connected to the metal center, and $\boldsymbol{m} \in [0, 1]^N$ encodes whether which atoms are part of the context. Group and connectivity information, i.e. $\boldsymbol{g}$ and $\boldsymbol{p}$, allow the model to discern the different connectivity patterns and differentiate between TS and non-TS, whereas structure-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 $\mathcal{M}$.

As a concrete example, a TMC such as depicted in Fig. 1 would feature 5 groups: a center (always set to 1), two A, one B, and one C ligands. As all considered ligands are monodentate, only the atom binding to the metal center would be considered as proximal in each of the groups except the group corresponding to the center. In the associated TS, $H_2$ would be added to the group containing the metal center, and both H atoms would be considered proximal.

**Conditional Equivariant Diffusion Model** We formulate a conditional probability distribution $p_\theta(\boldsymbol{r}^{\notin\mathcal{M}}, \boldsymbol{h}^{\notin\mathcal{M}} | \boldsymbol{r}^{\in\mathcal{M}}, \boldsymbol{h}^{\in\mathcal{M}}, \boldsymbol{c})$, that we parameterize using a conditional equivariant denoising network $\boldsymbol{\varepsilon}_\theta(\boldsymbol{z}_t, t, \boldsymbol{c})$, with $\boldsymbol{z}_t$ denoting the noisy version of $\boldsymbol{x}$ at time $t$. We employ a diffusion process and an architecture similar to that of OM-DIFF (Cornet et al., 2024a), with the only difference that the noisy atomic features $\boldsymbol{z}_t^{(h)}$ 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. 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.

**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 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 pair $(\boldsymbol{r}, \boldsymbol{h})$ appear thus multiple times, with different variations of $\boldsymbol{c}$. That way, $\boldsymbol{\varepsilon}_\theta$ is trained to denoise structures with varying contexts. We note that, depending on the downstream task, $\boldsymbol{m}$ can be constructed differently, *e.g.* at an atom-level instead of ligand-level.

**Conditional sampling of transition states (TSs)** During training, the model sees different coordination patterns, a mixture of isolated catalysts and TS structures, and varying contexts. At sampling time, the model is prompted with the desired conditional information, $\boldsymbol{c}_{\text{TS}}$, to generate TSs with a specific coordination pattern on demand. Namely, we specify the Ir metal centre, two atoms representing $H_2$ belonging to the same group as the center, where each H is proximal. Finally, we create $4$ groups of monodentate ligands. Partial generation around a known complex can be achieved by specifying $(\boldsymbol{r}^{\in\mathcal{M}}, \boldsymbol{h}^{\in\mathcal{M}})$ and $\boldsymbol{m}$ accordingly.

## 3 Experiments and Results

In what follows, we perform two experiments illustrating practical use cases. In Section 3.1, we sample novel complexes from scratch, yielding a broad coverage of the property space. In addition to evaluating the model, this can be useful to *e.g.* extend an existing database or to perform screening. In Section 3.2, we leverage the ability of the model to perform partial generation and concentrate the sampling around a known promising complex. We demonstrate that it effectively allows the property space to be searched locally.

### 3.1 Sampling from scratch

**Setup** We construct the necessary conditional information $\boldsymbol{c}_{\text{TS}}$ to generate a TS, and sample $10000$ complexes. The number of atoms composing the different ligands is drawn as follows: $N_{\text{A}} \sim \mathcal{U}([5, 40])$ $N_{\text{B}} \sim \mathcal{U}([1, 6])$ and $N_{\text{C}} \sim \mathcal{U}([2, 6])$. We provide a visualization of a handful of novel A ligands in Fig. 6.

**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 Appendix A.2. We evaluate the samples generated by each variant in terms of validity, uniqueness and novelty, as displayed in the top panel of Fig. 2. Details about evaluation can be found in Appendix C.1. We observe that augmenting the training data with other coordination patterns and non-TS geometries leads to a clear improvement of the generative capabilities of the model – an approximately $4$-fold improvement, increasing the success rate from around $12\%$ to nearly $50\%$.

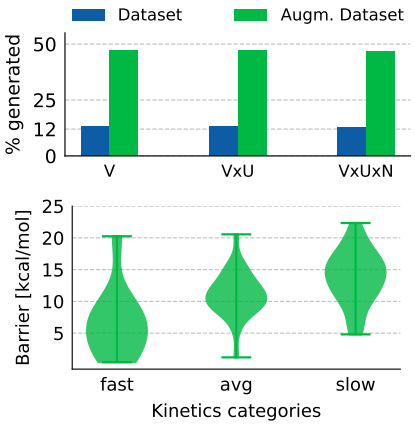


Figure 2: **From scratch experiment.** (**T**) Augmenting the dataset greatly increases Validity, Uniqueness and Novelty. (**B**) DFT energy barrier distribution for the kinetics categories estimated by the surrogate model.

**Data-driven screening** We then leverage a surrogate model (more details in Appendix D), and estimate the barrier energy for the V×U×N samples generated by the model trained on the augmented dataset. 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), average ($8.1$ kcal/mol $< \Delta E^{\ddagger}_{H_2} < 15.1$ kcal/mol), or slow ($\Delta E^{\ddagger}_{H_2} > 15.1$ kcal/mol). We then run DFT calculations (more details in Appendix E) to compute the true barrier of $\approx 400$ samples, of which $120$ were successful. We discuss the failure modes in more details in Appendix C.2. For each kinetics category, we display the corresponding energy distributions in the bottom panel of Fig. 2. We observe that the surrogate is capable of capturing the overall energy trend of the complexes generated by the diffusion model.

### 3.2 Searching in the neighborhood of a known complex

**Setup** We again provide $\boldsymbol{c}_{\text{TS}}$ to generate a TS structure, but we now also input a context set $\boldsymbol{x}^{\mathcal{M}}$ corresponding to complex 384 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 Fig. 4. 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 $N_{\text{A}} \sim \mathcal{U}([5, 40])$.

**Local search** We collect and run DFT calculations for around 200 randomly selected V×U×N samples. Among these, 50 were successful (more details in Appendix C.2). The resulting barrier energy distribution is provided in Fig. 3, where we can observe a clear shift towards lower barrier energies, compared to the initial dataset distribution. That shift can be explained by the low barrier energy (1.6 kcal/mol) of the seed complex. This 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.

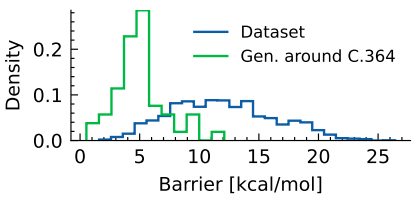


Figure 3: **Local search experiment.** Searching in the neighbourhood of complex 384, with a barrier energy of 1.6 kcal/mol, leads to a shift in the property distribution, as per DFT.

**Novel promising complexes** In Fig. 4, 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.

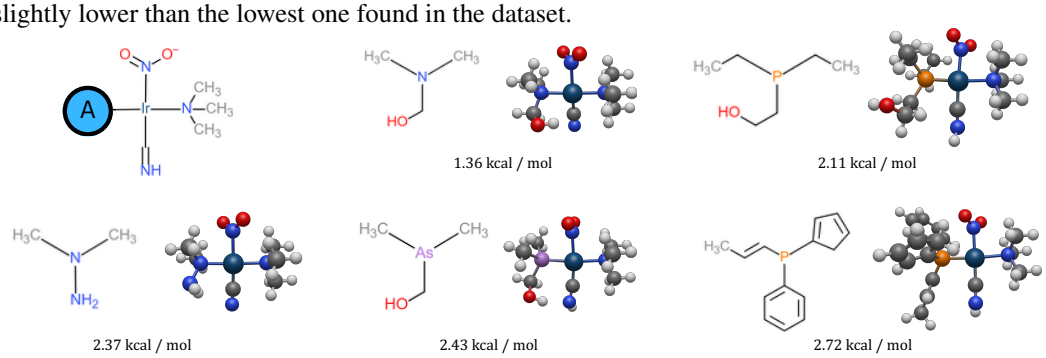


Figure 4: **Novel complexes obtained by redesigning one of the A ligands in complex** 384 extracted from the Vaska's dataset (Friederich et al., 2020). The context $\boldsymbol{x}^{\mathcal{M}}$ provided to the model is schematically depicted in the upper left corner. Out of the 50 calculations, these 5 complexes featured the lowest barrier energies. Each complex is presented with the designed A ligand.

## 4 Conclusion

We presented a conditional equivariant diffusion model that leverages 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 drastic improvement in the generative capabilities of the model, while maintaining full controllability at test time. We then showed that the model can effectively be combined with a surrogate model to perform screening in a fully data-driven fashion. Finally, we demonstrated that the model 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 while inducing a desired shift in the property distribution.

Avenues for future work are numerous. In terms of modelling, more advanced conditioning mechanisms (Wu et al., 2024; Denker et al., 2024), including bond information (Le et al., 2024), or more recent generative frameworks for geometric graphs (Irwin et al., 2024; Cornet et al., 2024b) are all promising directions.

## Acknowledgments and Disclosure of Funding

The authors acknowledge financial support from the Independent Research Fund Denmark with project DELIGHT (Grant No. 0217-00326B), and project TeraBatt (Grant No. 2035-00232B).

## References


Pascal Friederich, Gabriel dos Passos Gomes, Riccardo De Bin, Alán Aspuru-Guzik, and David Balcells. Machine learning dihydrogen activation in the chemical space surrounding vaska's complex. *Chemical Science*, 11(18):4584–4601, 2020.

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.

Ilia Igashov, Hannes Stärk, Clément Vignac, Arne Schneuing, Victor Garcia Satorras, Pascal Frossard, Max Welling, Michael Bronstein, and Bruno Correia. Equivariant 3d-conditional diffusion model for molecular linker design. *Nature Machine Intelligence*, pages 1–11, 2024.

Chenru Duan, Yuanqi Du, Haojun Jia, and Heather J Kulik. Accurate transition state generation with an object-aware equivariant elementary reaction diffusion model. *Nature Computational Science*, 3(12):1045–1055, 2023.

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

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.

Aditya Nandy, Chenru Duan, Michael G Taylor, Fang Liu, Adam H Steeves, and Heather J Kulik. Computational discovery of transition-metal complexes: from high-throughput screening to machine learning. *Chemical reviews*, 121(16):9927–10000, 2021.

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.

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.

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.

Hongni Jin and Kenneth M Merz Jr. Liganddiff: de novo ligand design for 3d transition metal complexes with diffusion models. *Journal of Chemical Theory and Computation*, 2024a.

Hongni Jin and Kenneth M Merz Jr. Partial to total generation of 3d transition-metal complexes. *Journal of Chemical Theory and Computation*, 2024b.

François Cornet, Bardi Benediktsson, Bjarke Hastrup, Mikkel N. Schmidt, and Arghya Bhowmik. Om-diff: inverse-design of organometallic catalysts with guided equivariant denoising diffusion. *Digital Discovery*, pages –, 2024a. doi: 10.1039/D4DD00099D. URL http://dx.doi.org/10.1039/D4DD00099D.

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.

Luhuan Wu, Brian Trippe, Christian Naesseth, David Blei, and John P Cunningham. Practical and asymptotically exact conditional sampling in diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024.

Alexander Denker, Francisco Vargas, Shreyas Padhy, Kieran Didi, Simon Mathis, Vincent Dutordoir, Riccardo Barbano, Emile Mathieu, Urszula Julia Komorowska, and Pietro Lio. Deft: Efficient finetuning of conditional diffusion models by learning the generalised $h$-transform. *arXiv preprint arXiv:2406.01781*, 2024.

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.

Ross Irwin, Alessandro Tibo, Jon Paul Janet, and Simon Olsson. Efficient 3d molecular generation with flow matching and scale optimal transport. In *ICML 2024 AI for Science Workshop*, 2024. URL https://openreview.net/forum?id=CxAjGjdkqu.

François RJ Cornet, Grigory Bartosh, Mikkel N Schmidt, and Christian A Naesseth. Equivariant neural diffusion for molecule generation. *Advances in Neural Information Processing Systems*, 37, December 2024b.

Efthymios I. Ioannidis, Terry Z. H. Gani, and Heather J. Kulik. molsimplify: A toolkit for automating discovery in inorganic chemistry. *Journal of Computational Chemistry*, 37(22):2106–2117, 2016. ISSN 1096-987X. doi: 10.1002/jcc.24437. URL http://dx.doi.org/10.1002/jcc.24437.

Aditya Nandy, Chenru Duan, Jon Paul Janet, Stefan Gugler, and Heather J. Kulik. Strategies and software for machine learning accelerated discovery in transition metal chemistry. *Industrial & Engineering Chemistry Research*, 57(42):13973–13986, 2018. ISSN 0888-5885. doi: 10.1021/acs.iecr.8b04015. URL https://doi.org/10.1021/acs.iecr.8b04015.

Greg Landrum. *RDKit: Open-source cheminformatics*, 2024. URL https://www.rdkit.org/.

John P. Perdew, Kieron Burke, and Matthias Ernzerhof. Generalized Gradient Approximation Made Simple. *Physical Review Letters*, 77(18):3865–3868, 1996. doi: 10.1103/PhysRevLett.77.3865.

Andreas Schäfer, Hans Horn, and Reinhart Ahlrichs. Fully Optimized Contracted Gaussian Basis Sets for Atoms Li to Kr. *Journal of Chemical Physics*, 97(4):2571–2577, 1992. doi: 10.1063/1.463096.

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. *Journal of Chemical Physics*, 132(15):154104, 2010. doi: 10.1063/1.3382344.

# A Dataset and data representation

## A.1 Vaska's complex dataset

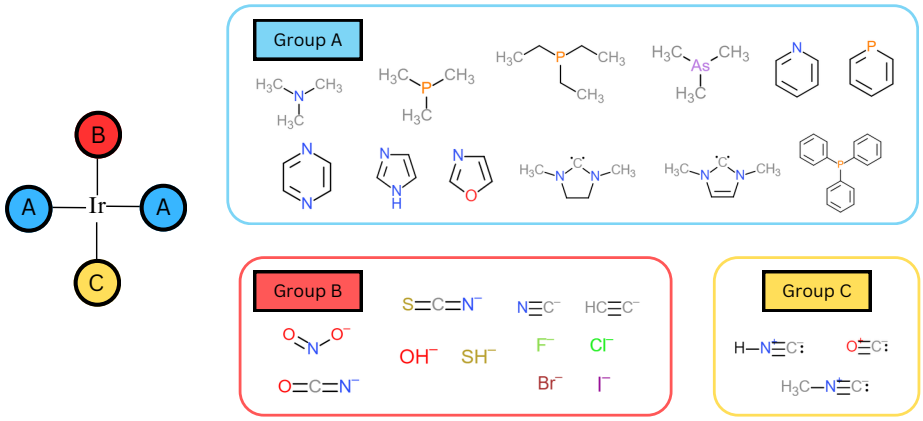


Figure 5: Ligands (26 in total) used to construct the initial dataset (Friederich et al., 2020). The ligands are grouped by type.

## A.2 Preparation of the augmented dataset

We leverage the updated TMQM dataset (Balcells and Skjelstad, 2020) to obtain additional training data. We extract linear, square-planar and octohedral complexes. We infer the coordination geometry using `MolSimplify` (Ioannidis et al., 2016; Nandy et al., 2018), and limit ourselves to complexes that only feature monodentate ligands and at most 100 atoms in total. This results in 12738 additional complexes. We also append the relaxed catalyst geometries from the original dataset (Friederich et al., 2020) to the training data. In total, we do have 15294 data points in the augmented dataset, of which around 2000 are TS geometries.

# B Generated ligands

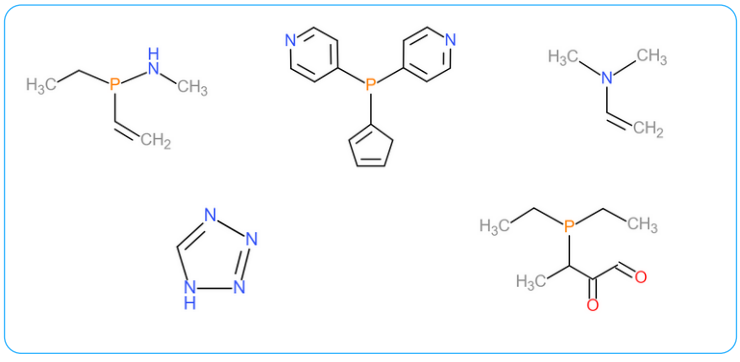

Figure 6: Excerpt of novel A ligands obtained in the generation from scratch experiment.

# C Evaluation details

## C.1 Checks

**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 Fig. 5. After bond inference, the resulting `Mol` should contains $4$ fragments, and `DetectChemistryProblems()` should return an empty list.

**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.

## C.2 Failure modes

While we only performed calculations on samples that were deemed valid (and novel) by the filters introduced Appendix C.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 $\approx 75\%$ 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 hours. 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.

## D 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 Fig. 7 and Table 1.

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

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

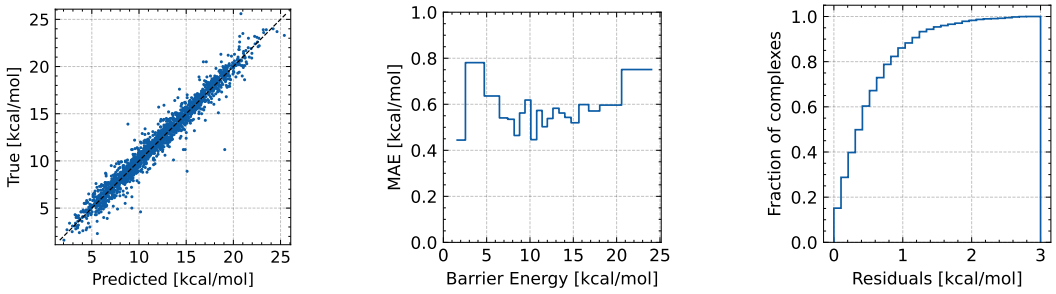


Figure 7: **Error diagnostic of the surrogate model.** (**L**) Parity plot. (**M**) MAE across the property space. (**R**) Cumulative distribution of the residuals.

## E DFT setup

A DFT protocol is set up to obtain the relaxed geometry of the catalyst, search for the transition state 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 transition state generated by the model, remove the activated hydrogens from the metal centre 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 with the PBE functional (Perdew et al., 1996). Optimizations were performed with the def2-SVP basis set (Schäfer et al., 1992). Effective Core Potentials (ECPs) 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 transition state search was carried out at the same level of theory. 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 transition state search. Frequency calculations were carried out to confirm that a proper transition state was obtained.

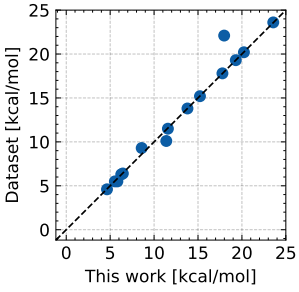


Figure 8: Plot showing our DFT setup yields the same values as the original one.