Directional Multiobjective Optimization of Metal Complexes at the Billion-Scale with the tmQMg-L Dataset and PL-MOGA Algorithm

Hannes Kneiding,[†] Ainara Nova,[†,‡] David Balcells[†,*]

[†]*Hylleraas Centre for Quantum Molecular Sciences, Department of Chemistry, University of Oslo, P.O. Box 1033, Blindern, 0315 Oslo, Norway;* [‡]*Centre for Materials Science and Nanotechnology, Department of Chemistry, University of Oslo, N-0315 Oslo, Norway*

E-mail: david.balcells@kjemi.uio.no

Abstract

Transition metal complexes (TMCs) play a key role in several areas of high interest, including medicinal chemistry, renewable energies, and nanoporous materials. The development of TMCs enabling these technologies remains challenged by the need to optimize multiple properties within very large chemical spaces, in which the thirty transition metals can be combined with a virtually infinite number of ligands. In this work, we provide the open tmQMg-L dataset including 30K TMC ligands, which combines large chemical diversity with synthesizability. The charge and metal-coordination mode of the ligands were robustly defined with a novel algorithm based on graph and natural bond orbital theories. The tmQMg-L dataset was leveraged in the automated generation of 1.37M TMCs resulting from all possible combinations between a square planar palladium(II) scaffold and a pool of 50 different ligands. This TMC space was used to benchmark a multiobjective genetic algorithm (MOGA) that optimized two properties over a Pareto front; namely the polarizability (α) and the HOMO-LUMO gap (ϵ). The MOGA evolved 130 TMC hits with maximal (α, ϵ) values in a way that could be easily rationalized by analyzing the nature of the ligands selected. Instead of the traditional mutation and crossover of fragments within a single ligand, this MOGA implemented full-ligand genetic operations acting on all coordination sites, maximizing chemical diversity. Further, we extended this MOGA algorithm with the Pareto-Lighthouse functionality (PL-MOGA), which allows for controlling both the aim and scope of the multiobjective optimization over the Pareto front. In explicit spaces containing billions of TMCs, the PL-MOGA enabled the explainable generation of thousands of novel and highly diverse TMC hits. We believe that the combined use of the tmQMg-L dataset and PL-MOGA algorithm will facilitate the discovery of TMCs with optimal properties within untapped chemical spaces.

Introduction

Transition metal complexes (TMCs) are chemical compounds of high interest due to their crucial role in diverse technologies, including medicinal chemistry,[1] catalysis,[2] electronic devices,[3] renewable energies,[4] and nanoporous materials.[5] From left to right, Figure 1 shows examples of chemotherapy drugs and cross-coupling catalysts based on square planar TMCs, as well as TMC alkyls used to produce semiconducting films, chelates promoting water splitting into oxygen and hydrogen, and oxides used as building blocks in MOFs. Current societal challenges, like the pandemics and the energy and climate crises, require accelerating the further development of TMC-based technologies. Computational modeling has proven successful in the design of new TMCs,[6–8] though this approach is often limited to small modifications of specific and previously known systems. Alternatively, the TMC chemical space can be systematically explored to find optimal compounds,[9,10] though its size is beyond the capabilities of the theoretical and experimental screening techniques currently available; for example, 10K different ligands can yield $> 10^{15}$ unique TMCs.

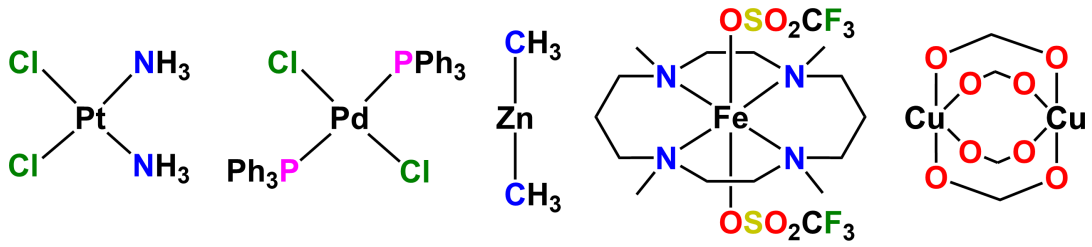

Figure 1: Examples of technology-relevant TMCs.

Despite the vastness of the chemical space, it is possible to cut large slices of it for their data-driven exploration with machine learning (ML) methods.[11] This approach has been successfully applied to catalysis[12,13] with TMCs,[14,15] including hydrogen activation,[16] C-H oxidation,[17] and C-O cleavage[18] reactions, using methods similar to those leveraged in organic chemistry,[19] drug discovery[20] and materials science,[21] in which the spaces explored are formulated explicitly. These spaces can contain thousands to millions of TMCs but their nature is often local, representing the neighborhood of one or few TMCs that are already

known. In principle, this issue can be tackled by formulating much larger implicit spaces if they can be explored with algorithms that do not need to generate all the TMCs within.

Genetic algorithms[22] (GAs) are the core method of evolutionary computation.[23,24] In chemistry and materials science,[25] GAs can be seen as generative models[26] mapping a desired target y into a set of features X defining a compound within an implicit chemical space. GAs thus act in an inverse design $y \leftarrow X$ fashion,[27] opposite to that of conventional ML predictive models. Despite some pitfalls, like the need for sampling many solutions to find the optima, GAs can match the efficiency of more complex ML methods[26,28] or be used to augment them.[29–31] GAs are particularly suitable for systems that, like TMCs, can be expressed as fragment combinations.[32] Figure 2 shows an example in which a ligand is evolved by mutation and crossover operations modifying and combining its fragments, respectively, to define new TMC generations in an iterative manner. A critical component of this setup is the use of ligand libraries, which, in the field of TMC chemoinformatics, are in general small in either size or diversity. For example, the *ligand knowledge bases*[33,34] and the *octahedral homoleptic ligand database*[35] are both diverse but their size is limited to a few thousand entries. In contrast, the *kraken* platform[36] allows for generating millions of ligands but all of them are monodentate phosphines.

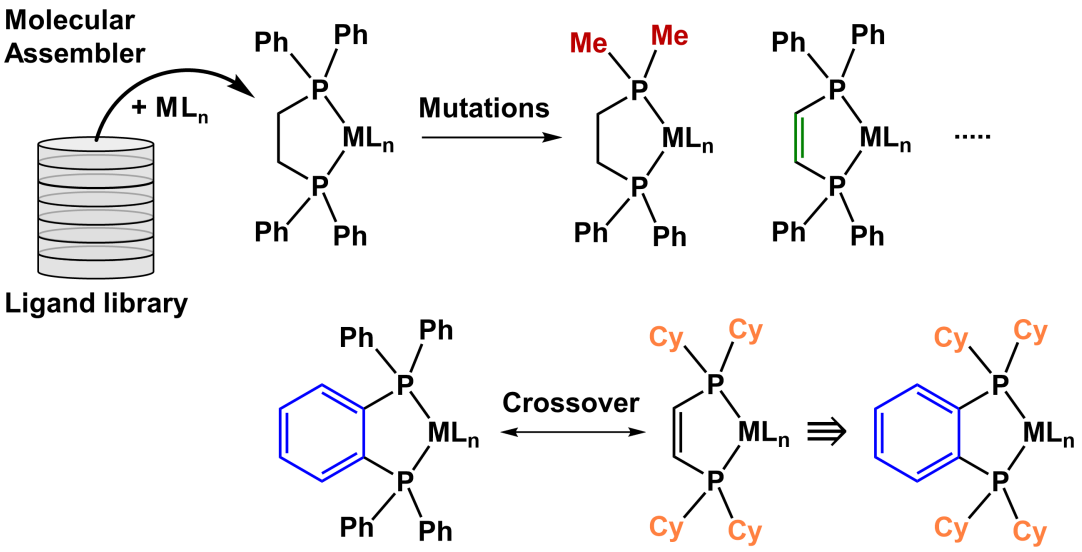

Figure 2: Mutation and crossover operations in a GA acting on a single TMC ligand.

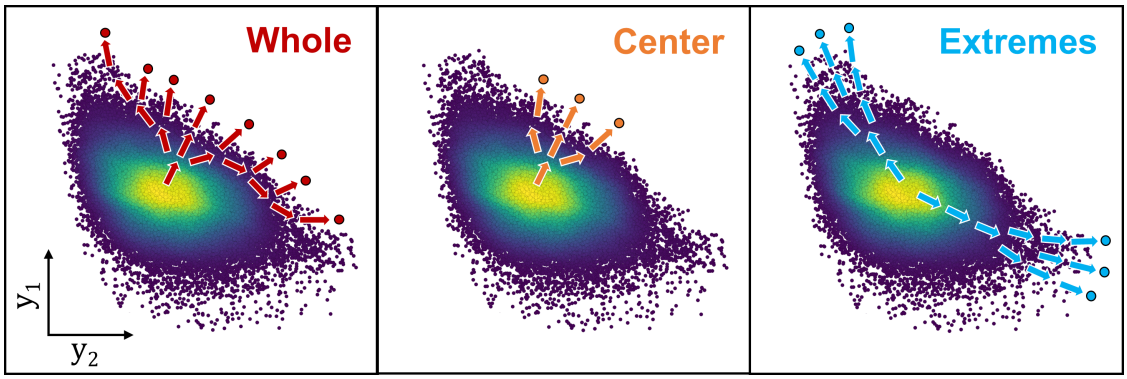

Figure 3: Directional multiobjective optimization.

Another key component in GAs for TMCs is the automated generation of guess geometries. The *molSimplify*,[37] *DENOPTIM*,[38] *AARON*,[39] *MolAssembler*,[40] and other[41] programs can assemble ligands around a metal center, producing geometries with the quality needed for quantum mechanics (QM) calculations. However, this requires a ligand library in which, for all entries, charge and metal coordination are well-defined. Coordination is easily determined if the TMC graph can be derived from its formula or geometry, though this is not trivial[42] due to the complexity of metal–ligand bonds. Defining the charge is also difficult and it is a major factor limiting the construction of libraries from experimental sources like the Cambridge Structural Database (CSD).[43] Previous approaches have focused on screening homoleptic TMCs[44] and assigning metal oxidation states.[45] Alternatively, well-defined libraries can be made *ad hoc in silico*[46] but this can compromise synthesizability.[47,48]

Once the chemical space is set, a fitness function is defined to rank the TMCs evolved by the GA. The fitness is optimized over multiple generations and it should therefore be a computationally inexpensive $y = f(X)$ function, where X specifies the TMC geometry and y its fitness. f can be a QSAR model[49] or a QM method,[50] and y can refer to one or several target properties, depending on whether the GA optimization is single- or multi-objective. The latter is intrinsically more challenging, especially when it involves uncorrelated molecular properties forming a Pareto front and one wants to direct the optimization to a specific region (Figure 3). Multiobjective optimization with genetic algorithms (MOGA) has been

implemented with different methods,[51–53] including the *NaviCatGA* platform,[54] without requiring the explicit spaces used in other ML approaches.[55,56] GAs have been applied to the *de novo* design of catalysts for the olefin metathesis,[49] oxidative C-C cleavage,[54] and Baylis–Hillman[57] reactions. With TMCs, the genetic operations have focused on a single ligand within a previously known system (Figure 2). This approach is efficient at exploring the chemical space but it can also compromise synthesizability or halt the GA into local minima.

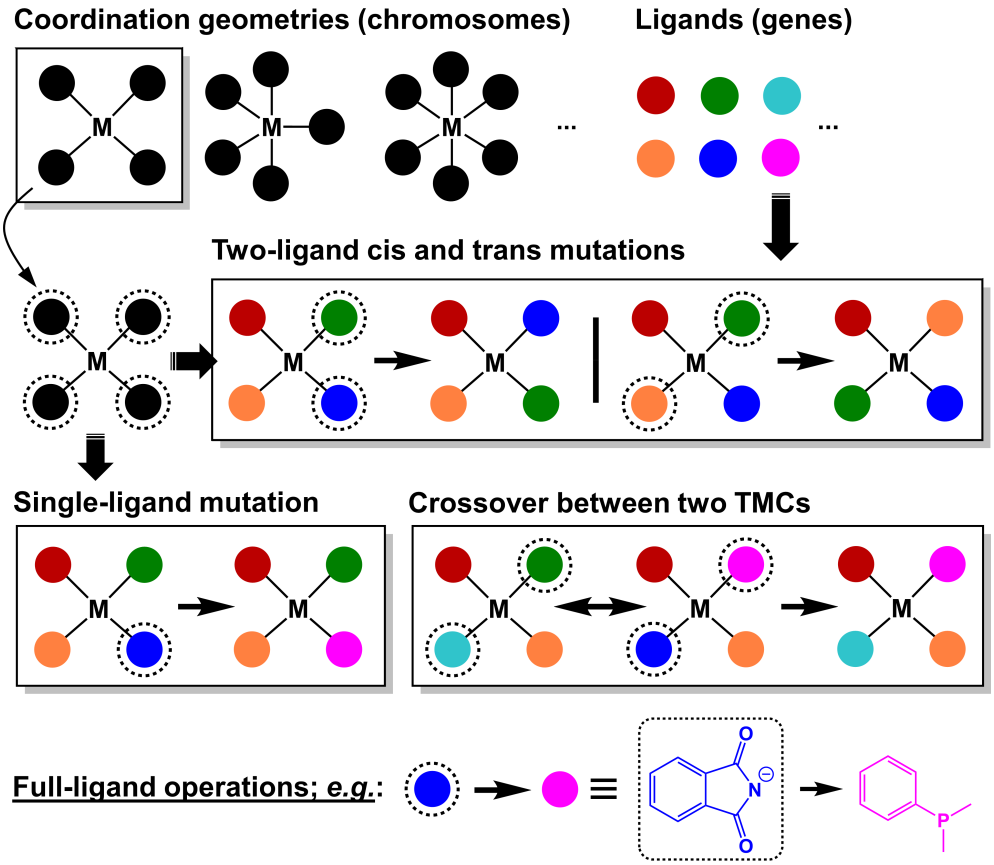

Figure 4: Multiligand MOGA in a square planar TMC. M = metal.

In this work, we report the 30K tmQMg-L library, which provides an extensive, diverse, and synthesizable set of TMC ligands extracted from the CSD. This library was compiled as a dataset including metal-binding geometries and electronic and steric information, in an open format allowing for automated workflows. Ligand charges and metal coordination modes were robustly defined with a new algorithm based on NBO and graph analyses.

tmQMg-L was leveraged in a novel MOGA that, instead of doing local modifications of a single ligand (Figure 2), does full-ligand genetic operations over all coordination sites of any given TMC scaffold, considering the coordination geometry and its isomerism, as shown in Figure 4 for the square planar TMC space. This MOGA thus has the ability to evolve unprecedented combinations of known ligands within diverse and vast combinatorial spaces. We also developed the Pareto-Lighthouse MOGA (PL-MOGA) which, through a simple and intuitive selection of scaling factors, allows for fine-tuning the aim of the optimization over the Pareto front, controlling not only its direction, as shown in Figure 3, but also its scope. After benchmarking this evolutionary method in the multiobjective optimization of the polarizability (α) and the HOMO-LUMO gap (ϵ) in a space of 1.37 million square planar Pd(II) TMCs, we explored implicit spaces containing billions of TMCs. The algorithm evolved TMC hits maximizing one or both target properties in an explainable manner, and keeping chemical diversity and originality high with a low computational cost.

The tmQMg-L ligand dataset

tmQMg-L is a dataset belonging to the transition metal Quantum Mechanics (tmQM) series.[42,58] It provides 29,764 (30K) ligands extracted from the tmQMg dataset, which contains the geometry and electronic structure properties of 60,799 TMCs. All these complexes are present in the CSD, and, therefore, all ligands in tmQMg-L exist in at least one TMC for which the experimental crystal structure and synthesis have been reported. This can be a useful feature for enforcing synthesizability in generative models.[47] Further derivatization of the ligands through decoration with functional groups is also possible using the data provided. The tmQMg-L dataset is diverse in terms of the ligand charge, metal-coordinating elements, and metal-coordination modes, bulkiness, and ligand field strength. Figure 5 shows a random sample of 28 ligands illustrating the diversity of the dataset, with examples in these 12 different categories: phosphines, alkyls, carbenes, chelating amines, allyls, olefins, arenes, carboxylates, amidos, arsinines, halides, and pincers.

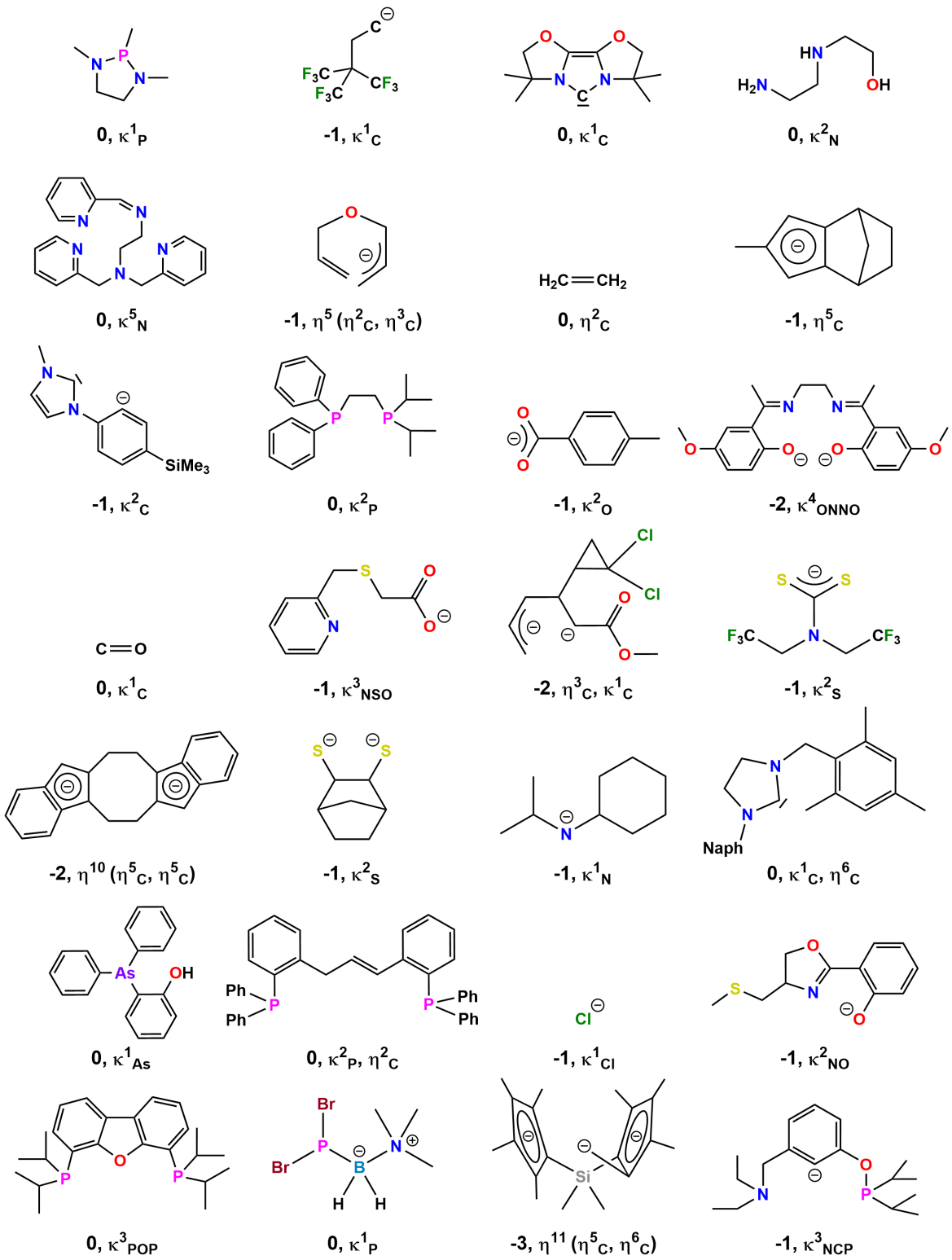

Figure 5: A random selection of 28 ligands illustrating the diversity of the tmQMg-L dataset. Each ligand is labeled with its charge, followed by the denticity (κ) and hapticity (η), including both order and elements involved, as determined by the protocol shown in Figure 6.

The TMC dataset from which the ligands were extracted, tmQMg, is also a graph dataset; *i.e.*, each TMC is defined by an undirected natural quantum graph (u-NatQG),[42] which is a representation based on natural bond orbital (NBO) theory yielding molecular graphs similar to the skeletal formulas used by organometallic chemists. From the u-NatQG graphs, the extraction of the tmQMg-L ligands is trivial; after removing the metal node, the resulting disconnected subgraphs are immediately identified as the ligands (Figure 6). Further, the node and edge indices defining the topology of the graphs allowed for describing the ligand coordination, including the atomic numbers defining the metal-bound elements (E) and their coordination in terms of both order (n) and mode (κ or η); *i.e.* κ^n_E denticity, for n non-contiguous atoms, and η^n_E hapticity for n contiguous atoms.

Whereas extracting the tmQMg-L ligands was straightforward, assigning their charge was challenging. In a recent account, Ess and co-workers developed an ML approach to ligand charge assignment based on QM features.[59] After experimenting with different data sources and algorithms, we found that the most reliable approach was to use the Lewis structures available from the NBO data of the tmQMg dataset[42] (Figure 6). By counting the number of bonding (N_{BD}) and lone (N_{LP}) electron pairs relative to the number of valence electrons (N_{V_e}), we could derive the formal charges of all atoms, which, upon summation, yielded the overall charge of the ligand (q_L in equation 1). The electron pairs of the metal–ligand bonds (N^{M-L}_{BD}) were counted as filling the valences of the metal-bound atoms (*i.e.* ionic electron counting).

$$q_L = \sum_{i}^{atoms \in L} -(N_{BD} + 2N_{LP} + 2N^{M-L}_{BD} - N_{V_e}) \quad (1)$$

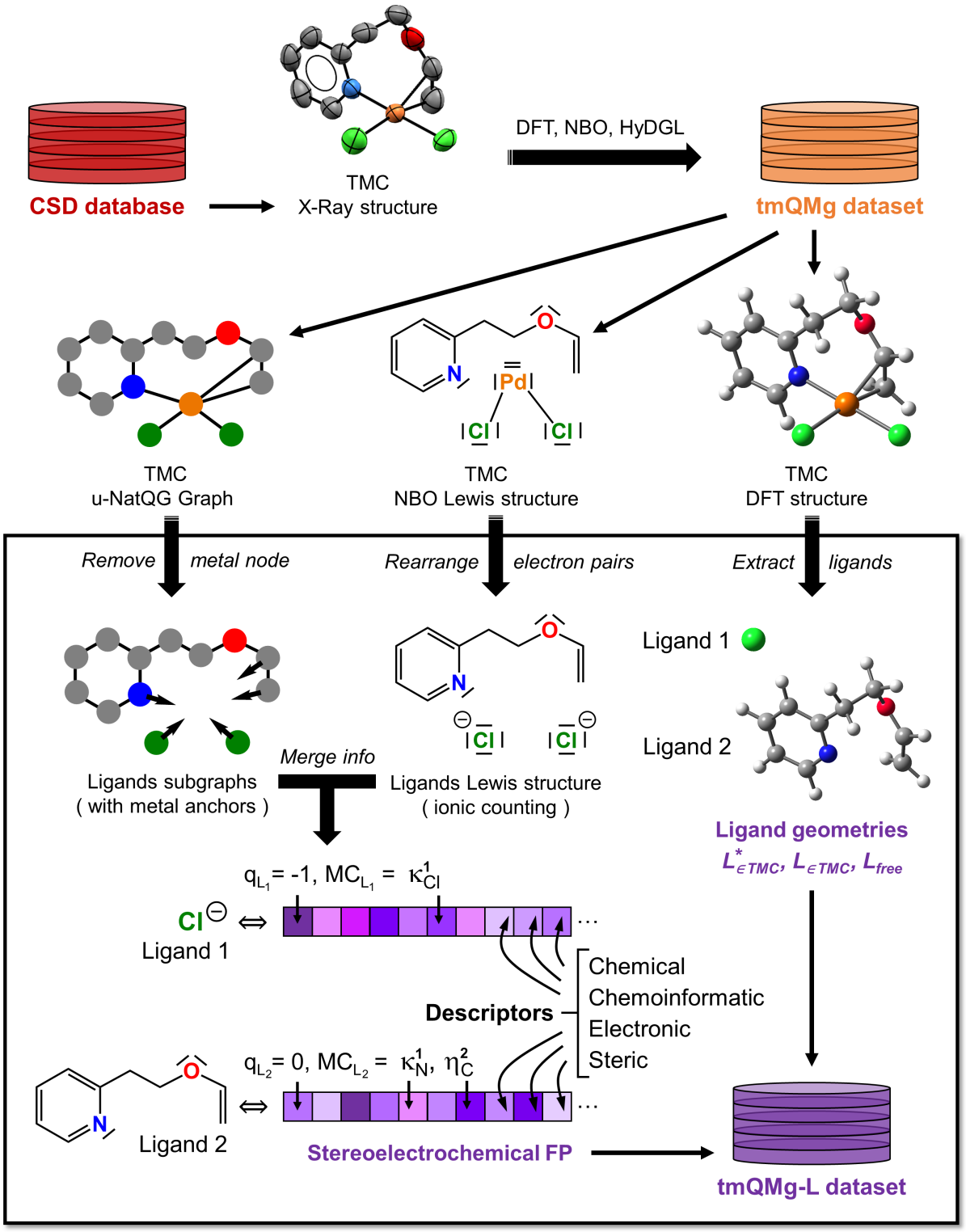

Figure 6: Derivation of the tmQMg-L ligand dataset. L = ligand; TMC = transition metal complex; NBO = natural bond orbital; HyDGL = Hylleraas deep graph learning program; q_L = ligand charge; MC = metal coordination; FP = fingerprint; $L^*_{\in TMC}$ = in-TMC L geometry yielding the lowest energy; $L_{\in TMC}$ = collection of in-TMC L geometries; L_{free} = optimized metal-free L geometry.

In order to assess the robustness of our protocol in assigning charges and coordination modes, we manually inspected 500 random ligands from tmQMg-L, which were selected by considering their size in atoms, charge, metal-bound elements, and metal-coordination mode. With the aim of maximizing the diversity and representativity of the selection, the different categories associated with these variables (*e.g.* κ and η, for the metal coordination mode) were included in the same ratio observed in the whole dataset. For all 500 ligands, the charge assigned to the ligand was consistent with a singlet spin multiplicity. The success rate in the assignment of the charge was 95%; *i.e.* for 475 of the 500 ligands, the charge was the most commonly observed in TMCs (*e.g.* alkyl- and aryl-phosphines were assigned q = 0 instead of –2 or +2, which would also be spin-consistent but rare). The other 25 ligands were assigned spin-consistent but unusual charges. Further, the indices of the metal-bound atoms were found to be correct for all 500 ligands, whereas the metal coordination mode was correctly defined for 485 of the 500 ligands assessed (97%). Considering the assignment of both the ligand charge and the coordination mode yielded an overall success of 92%, reflecting the robustness of this approach.

It should be noted that, depending on several factors, like the synthesis of the TMCs and the nature of their metal centers, there are ligands for which the charge can take different values and the coordination to the metal can occur in different modes. For example, regarding the charge, the O_2 ligand can be either neutral (*i.e.* dioxygen ligand) or anionic (*i.e.* superoxo O_2^{-1} and peroxo O_2^{-2}). Regarding the coordination mode, whereas the superoxo often coordinates in κ^1 fashion, the peroxo prefers the η^2. Further, these two factors are not necessarily related. For example, aliphatic ligands with a single carboxylate functional group will have a unique charge of -1e, and yet they may coordinate to the metal in either κ^1 or κ^2 fashion. For this type of ligands, the tmQMg-L dataset may contain either one or multiple charge and coordination variants. Using the provided fingerprints (*vide infra*), these ligands can be easily distinguished and found in the dataset and, if needed for a given task, their charge and coordination mode can be further diversified.

Table 1: Systematic list of all features included in the stereoelectrochemical fingerprints of the tmQMg-L ligand dataset. NBO = Natural Bond Orbital theory; NatQG = Natural Quantum Graph;[42] DFT1 = PBE/def2SVP optimization; DFT2 = PBE0/def2TZVP single-point.

Chemical properties. General				
Label	Definition	Units	Method/Software	Structure(s)
F	Chemical formula	—	Hill system	—
M	Molecular mass	Da	—	—
q_L	Molecular charge	e	NBO	—
N_A	Total number of atoms	—	—	—
$N_{A,E}$	Number of atoms by element	—	—	—
N_e	Number of electrons	—	—	—
Pop	Occurrences in tmQMg	—	—	$L_{\in TMC}$
Chemical properties. Coordination mode				
Label	Definition	Units	Method/Software	Structure
N_{MB}	Number of metal-bound atoms	—	NatQG	$L_{\in TMC}$
κ^n	Denticity order	—	NatQG	$L_{\in TMC}$
κ^n_E	κ^n by element	—	NatQG	$L_{\in TMC}$
η^n	Hapticity order	—	NatQG	$L_{\in TMC}$
η^n_E	η^n by element	—	NatQG	$L_{\in TMC}$
Cheminformatics descriptors				
$SMILES$	SMILES string	—	RDKit	—
MFP	Morgan fingerprints	—	RDKit	—
$logP$	Octanol/water partition coefficient	—	RDKit	—
$N_{R,Al}$	Num. of aliphatic rings	—	RDKit	—
$N_{R,Ar}$	Num. of aromatic rings	—	RDKit	—
$N_{R,Sat}$	Num. of saturated rings	—	RDKit	—
N_{RB}	Num. of rotatable bonds	—	RDKit	—
Electronic properties				
Label	Definition	Units	Method/Software	Structure
α	Polarizability	Bohr^3	DFT1	L_{free}
μ	Dipole moment	D	DFT1, DFT2	L_{free}, $L^*_{\in TMC}$
ν	Largest vibrational frequency	cm^{-1}	DFT1	L_{free}
ϵ	HOMO-LUMO gap	Ha	DFT1, DFT2	L_{free}, $L^*_{\in TMC}$
E^{MB}_{HOMO}	Metal-bound HOMO energy	Ha	DFT1, DFT2	L_{free}, $L^*_{\in TMC}$
S^{MB}_{HOMO}	Metal-bound HOMO symmetry	—	DFT1, DFT2	L_{free}, $L^*_{\in TMC}$
E^{MB}_{LUMO}	Metal-bound LUMO energy	Ha	DFT1, DFT2	L_{free}, $L^*_{\in TMC}$
S^{MB}_{LUMO}	Metal-bound LUMO symmetry	—	DFT1, DFT2	L_{free}, $L^*_{\in TMC}$
Steric properties				
Label	Definition	Units	Method/Software	Structure
V	Molecular volume	Å^3	DFT1, DFT2	L_{free}, $L^*_{\in TMC}$
$\frac{I_1}{I_3}, \frac{I_2}{I_3}$	Principal moments of inertia ratios[60]	—	DFT1, DFT2	L_{free}
Ec	Eccentricity	Å	RDKit	L_{free}, $L^*_{\in TMC}$
A^{SAS}	Solvent accessible surface area (SAS)[61,62]	Å^2	Morfeus	L_{free}, $L^*_{\in TMC}$
V^{SAS}	Volume within SAS[61,62]	Å^3	Morfeus	L_{free}, $L^*_{\in TMC}$
V_{Bur}	Buried volume[63]	%	Morfeus	$L^*_{\in TMC}$
θ°	Exact cone angle[64,65]	°	Morfeus	$L^*_{\in TMC}$
Ω	Solid angle[66,67]	sr	Morfeus	$L^*_{\in TMC}$
Θ	Solid cone angle[66,67]	°	Morfeus	$L^*_{\in TMC}$
G	G parameter[66,67]	—	Morfeus	$L^*_{\in TMC}$

Both the charge and the coordination mode of each ligand are included in an extensive stereoelectrochemical fingerprint provided with the dataset (Figure 6). The information included in the fingerprints allows for a fast search and systematic analysis of the tmQMg-L ligands, and their multiple features can be leveraged in machine learning models. These features include SMILES strings, molecular properties, orbital energies and symmetries, and diverse measures of shape and bulkiness. Table 1 gives a systematic list of all features included in the fingerprint.

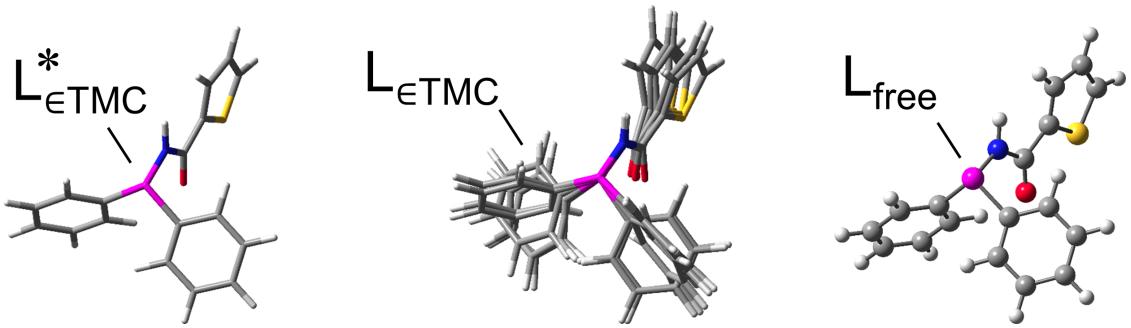

Figure 7: The $L^*_{\in TMC}$, $L_{\in TMC}$, and L_{free} geometry sets for an example phosphine ligand in tmQMg-L. Element color code: Violet (P), blue (N), grey (C), red (O), yellow (S), and white (H).

In addition to the fingerprint properties, tmQMg-L also provides geometric information in two distinct categories (Figures 6 and 7). One category corresponds to the structure of the ligand as it is within the DFT-optimized TMC from which it was extracted ($L_{\in TMC}$), whereas the other corresponds to the metal-free ligand (L_{free}). The $L_{\in TMC}$ category contains either one or multiple geometries, depending on whether the ligand was found in one or more TMCs of the original tmQMg dataset. When there are multiple geometries, the most stable one ($L^*_{\in TMC}$) is provided first, followed by the others ordered by increasing energy. These geometries do not differ by either charge or coordination mode, which would generate additional different entries in the ligand dataset, but rather structurally, in spatial regions far from the metal center (*e.g.* different conformations of aliphatic chains). The L_{free} structural category contains a single geometry, which is that of the fully optimized free ligand (*i.e.* not bound to the metal center), starting from $L^*_{\in TMC}$. This geometry was mainly computed for the sake of deriving other properties depending on energy derivatives (*e.g.* the largest

vibrational frequency). Table 1 shows how these structural categories relate to the features of the ligand fingerprint, and Figure 7 shows an example of these three sets for a phosphine ligand. For the generation of TMC geometries with automation software like *molSimplify*, we recommend using the $L^*_{\in TMC}$ geometry. All geometry optimization and single-point energy calculations were performed at the DFT(PBE/def2SVP) and DFT(PBE0/def2TZVP) levels of theory, respectively, in the closed-shell singlet state.

1.37M explicit space

The tmQMg-L ligand dataset can be leveraged in the automated construction of vast TMC spaces. With the aim of assessing this application, we explored the palladium square planar scaffold, which is very popular in the fields of metallodrugs and catalysis. For a Pd(II) center, restricting the ligand choice to monoanionic (n_a) and neutral (n_n) monodentate ligands, and the overall charge of the resulting complex to {-1, 0, 1}, the total number of unique TMCs (N) is given by

$$N = n_a^3 n_n + \frac{3}{2} n_a^2 n_n^2 + n_a n_n^3 + \frac{1}{2} n_a n_n \tag{2}$$

This expression accounts for the rotation invariances and cis/trans isomerism of the square planar coordination geometry for all possible $M(L)_4$, $M(L)_3(L')$, $M(L)_2(L')_2$, $M(L)_2(L')(L'')$, and M(L)(L')(L")(L"') formulations (Figure 8; see SI for a full derivation of Eq. 2). The 8418 monodentate ligands that are either neutral or monoanionic within tmQMg-L generate a massive chemical space of $1.26{\cdot}10^{15}$ TMCs. Further, Figure 8 shows that the 1 million and 1 billion marks in the size of this particular TMC space are surpassed by using only ∼50 and ∼250 different ligands, respectively.

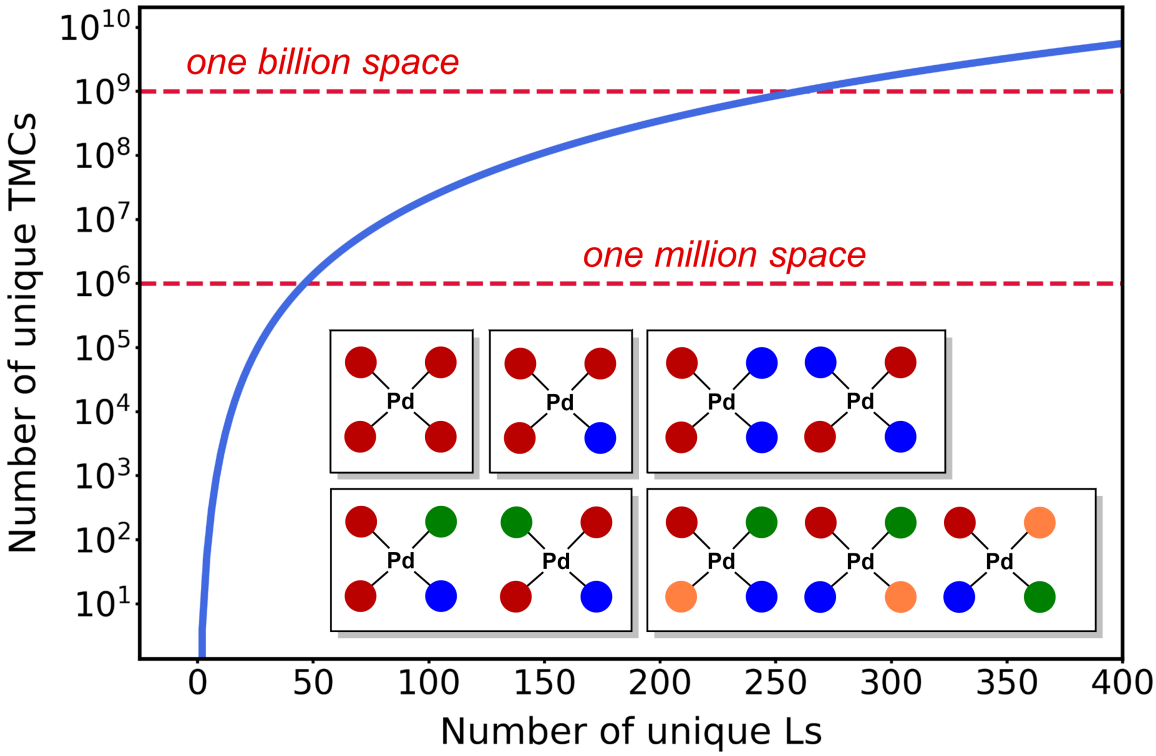

Figure 8: Chemical space size of a square planar palladium(II) scaffold as a function of the number of unique monodentate ligands that can be coordinated to it (Eq. 2). The charges of the ligands and the resulting TMCs were restricted to {-1, 0} and {-1, 0, +1}, respectively. The inset shows the different possible symmetries depending on the number of different ligands within a single TMC.

Under these ligand charge and metal-coordination constraints, we created a Pd(II) square planar TMC space using the 50 most popular ligands, half of them neutral and the other half monoanionic, and limiting their size to a maximum of 15 heavy atoms (Figure 9). Popularity hereby refers to the number of occurrences; *i.e.* the number of palladium TMCs within the tmQMg dataset containing a given ligand. Using *molSimplify* to automate the generation of the TMCs, these ligands yielded 1,367,485 (1.37M) geometries, which were fully optimized at the semiempirical GFN2-xTB level of theory. Figure 10 shows a scatter plot of their polarizability (α) and HOMO-LUMO gap (ϵ). Within this large chemical space, α and ϵ appear poorly correlated and distributed over wide ranges; *i.e.* ∼50-475 Bohr^3 and ∼0.15-4.15 eV, respectively.

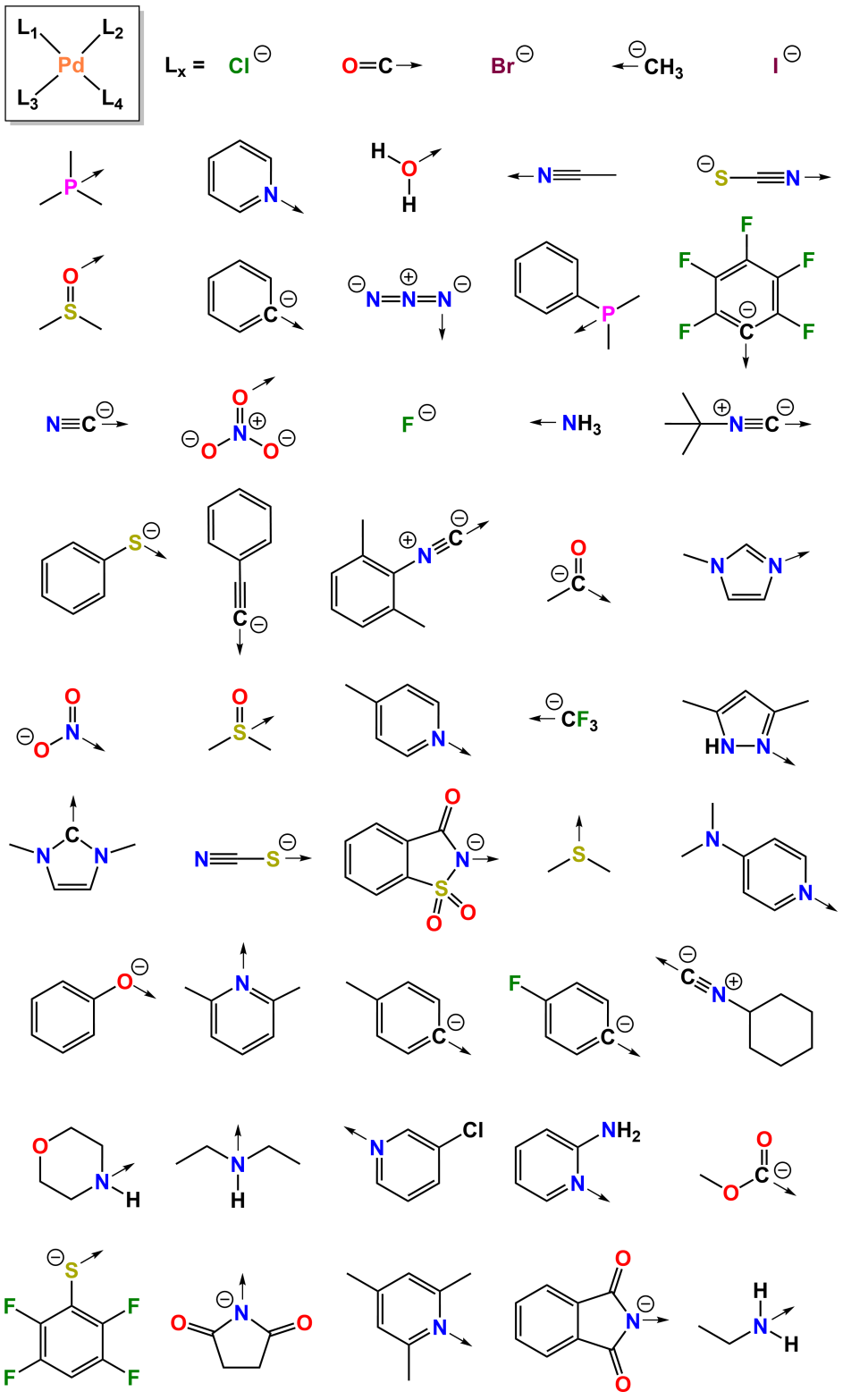

Figure 9: The 50 ligands used in the generation of the 1.37M chemical space ordered by popularity from top-left to bottom-right. In the polyatomic ligands, the arrow signals the metal-coordinating atom.

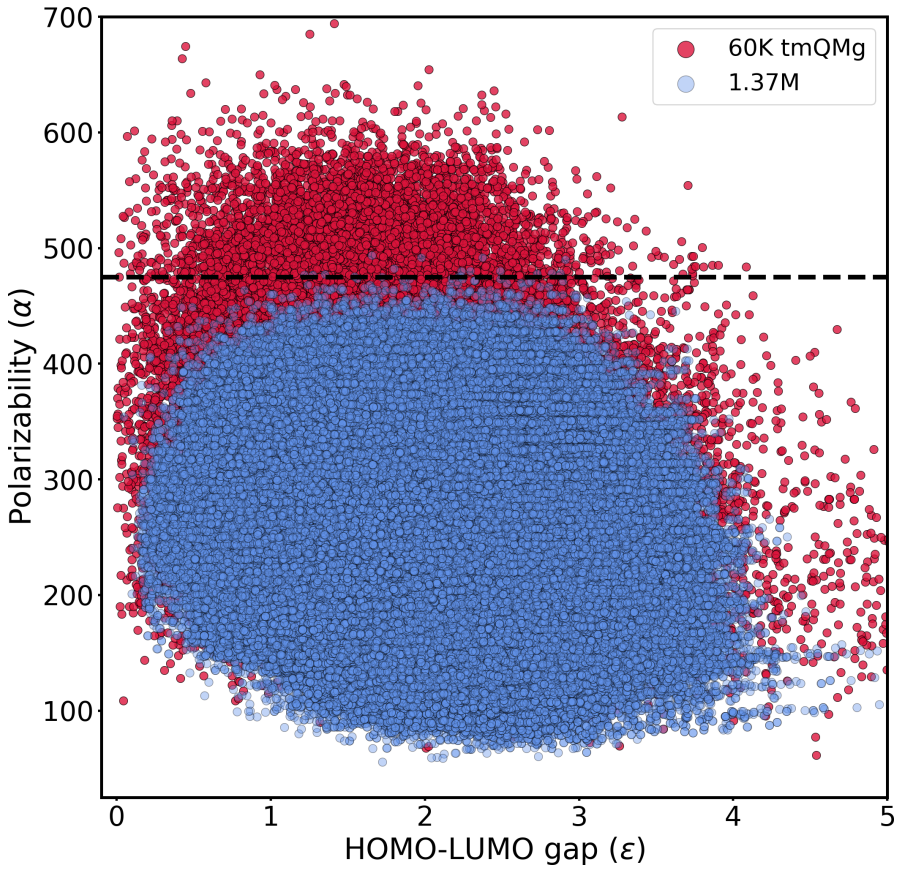

Figure 10: (α, ϵ) distribution at the GFN2-xTB level of theory for the tmQMg and 1.37M chemical spaces. The dashed line corresponds to $\alpha = 475$ Bohr^3. The entire 1.37M distribution plotted on top is covered by the 60K tmQMg underneath. The α and ϵ units are Bohr^3 and eV, respectively.

Figure 10 plots the 1.37M data against the tmQMg space, which contains 60K TMCs for which α and ϵ were computed at the same level of theory. There is a large overlap between both spaces except for $\alpha > 475$ Bohr^3, due to the size limit applied to the ligands used to build the 1.37M space. Interestingly, below this threshold, and thus for most of the (α,ϵ) range, this space covers the tmQMg. This may seem an expected observation because tmQMg is more than one order of magnitude smaller than the 1.37M space. However, whereas tmQMg is a CSD collection of TMCs combining the 30 transition metals with 30K different ligands, the 1.37M space is based on only one metal and 50 ligands. This shows that full combinatorial explosions with a limited number of metals and ligands can cover significant portions of the properties associated with the TMC space. In the next two sections, the (α,ϵ) data of the 1.37M space was used to benchmark the MOGA and PL-MOGA algorithms.

MOGA benchmark

Besides the overlap, Figure 10 also shows the poor correlation between α and ϵ, which, due to their relationship (*i.e.* large ϵ values limit the extent to which α can be maximized, and the other way around), form a Pareto front. The interplay between these two molecular properties is relevant in the field of drug discovery, since, ideally, a commercial active compound would maximize both α, for electrostatic and Van der Waals interactions with biomolecules, and ϵ, for stability against moisture, heat, or light. Given the interest in these two properties and the use of square planar metal complexes as chemotherapy drugs, we decided to tackle the Pareto front optimization of α and ϵ. In particular, we implemented a MOGA that used the genetic operations shown in Figure 4. The parent TMCs of each generation were selected with the probabilistic roulette-wheel method, after ranking them according to the non-dominating fronts of the populations. In contrast, the survivors were selected in a deterministic manner, ranking them according to the number of TMCs they are dominated by (see SI for further details).

The MOGA (α,ϵ) optimization was benchmarked in the explicit 1.37M space (Figure 10), for which the GFN2-xTB ground truth was known. We set the goal of finding 130 hits by exploring 13,000 TMC candidates (*i.e.* 0.01% and 1% of the entire space, respectively), which was implemented by evolving 100 generations with the genetic operations shown in Figure 4. The progress of the MOGA optimization is shown in Figure 11. After the first random generation, the TMCs evolved in the 10^{th} generation already appeared clustered over a wide (α,ϵ) region parallel to the Pareto front. The solution cluster quickly thinned and advanced and, by the 50^{th} generation it was mostly converged, suggesting that an exploration of a smaller space of solutions would also be efficient. In the last generation of hits, the 100^{th}, all TMCs are over the Pareto front and 18 were dominating points.

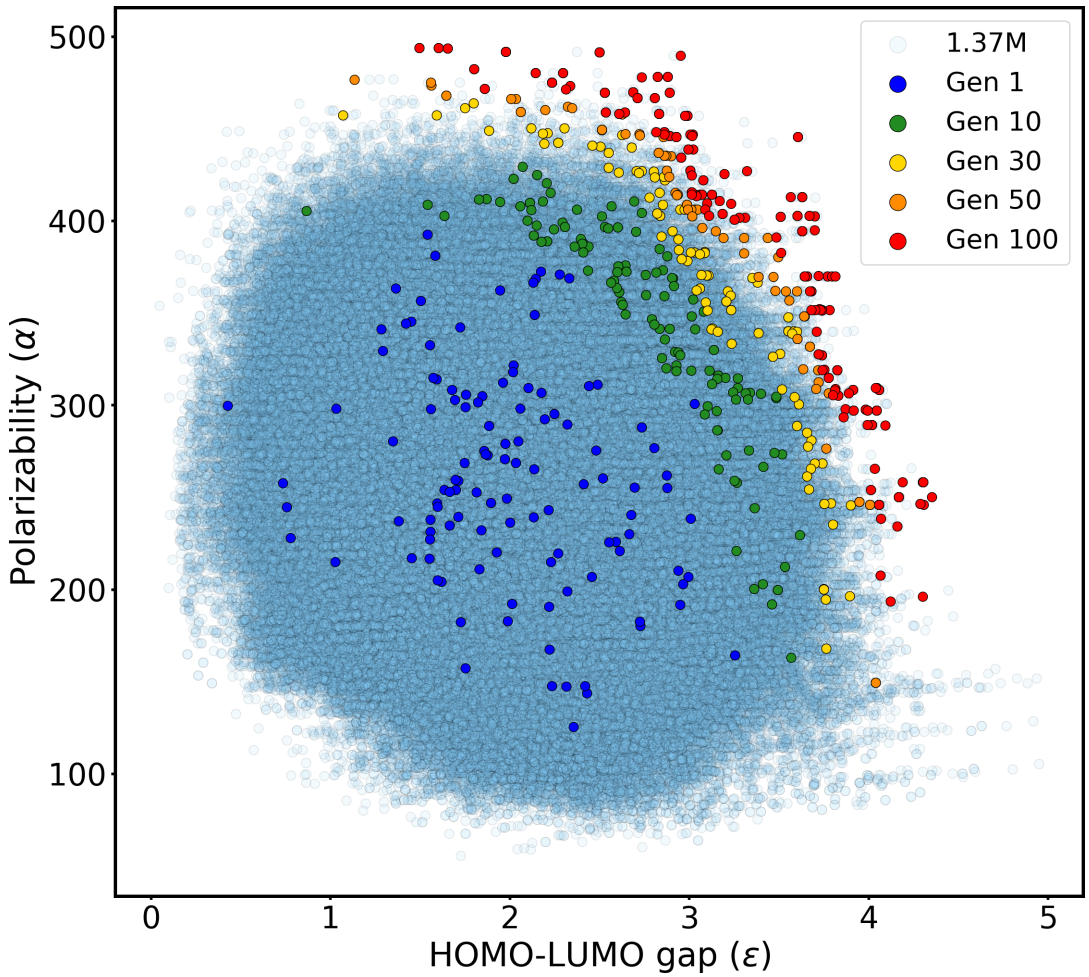

Figure 11: Multiobjective (α,ϵ) optimization in the 1.37M space with the MOGA algorithm. The α and ϵ units are Bohr^3 and eV, respectively.

In order to rationalize the progress of the MOGA toward the Pareto front, we plotted a histogram (Figure 12) showing the absolute and relative frequencies with which the different ligands were used. This plot shows that the MOGA enforced diversity by using all 50 ligands in the pool (Figure 9) over the 100 generations evolved. The 10 most popular ligands indicated a clear trend toward picking the ligands that maximize α with aromatic rings (*e.g.*, the P(Ph)(Me)$_2$ phosphine), ϵ with strong field coordinating moieties (*e.g.* the CN(Cy) isocyanide), and both α and ϵ by combining these features (*e.g.* $C_6F_5^-$ ligand).

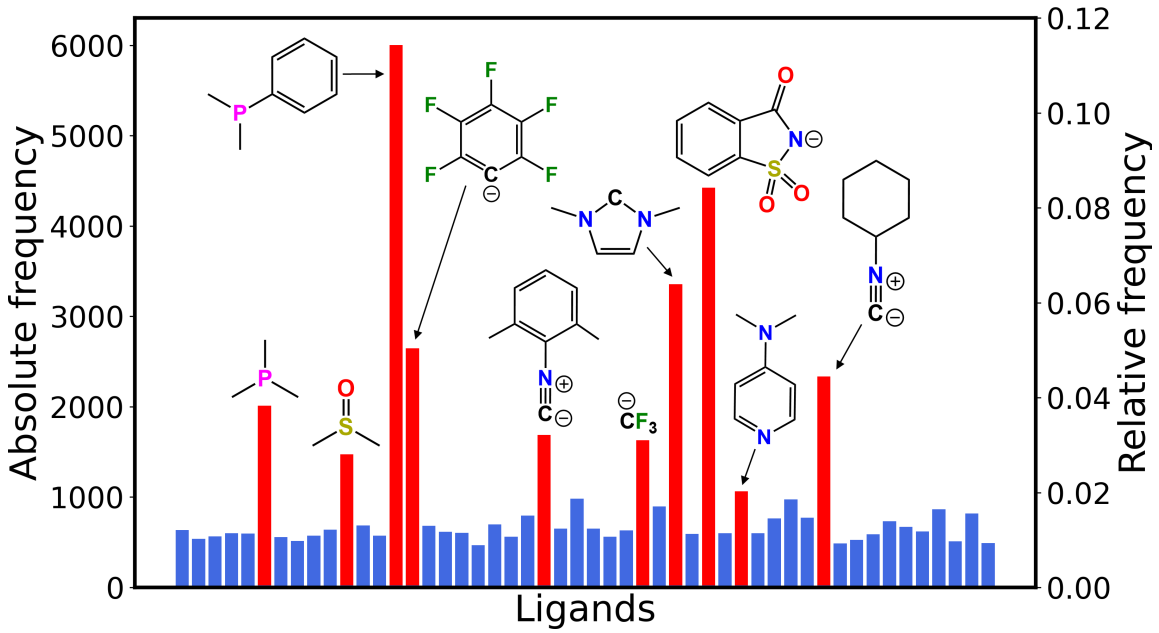

Figure 12: Ligand use absolute and relative frequencies in the (α,ϵ) Pareto front optimization. The red bars correspond to the ten most used ligands. From the left- to the right-hand sides, ligands are ordered by popularity as they are in Figure 9.

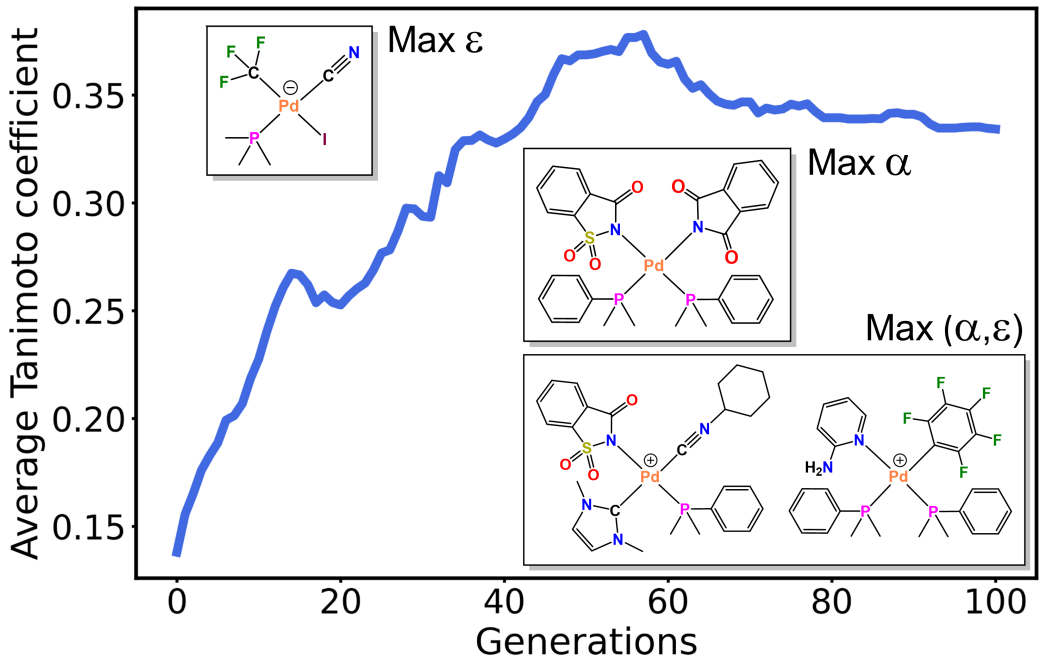

Figure 13: Average Tanimoto coefficient over the MOGA generations. The insets show random examples of TMC hits maximizing the polarizability (α), the HOMO-LUMO gap (ϵ), or both.

GAs tend to converge into local minima, limiting the diversity of the elite generations. However, this is not a significant issue in the present study, as shown by Figure 11, in which the hits appear scattered over the whole Pareto front, and Figure 12, in which at least 10 different ligands stand over the base frequencies. Further, of the 50 ligands of the pool (Figure 9), nearly half of them, 24, are present in the TMC hits of the last generation. In order to assess chemical diversity, we plotted the change in the average Tanimoto coefficient (TC) over the MOGA run (Figure 13), based on the concatenation of the four ligand SMILES strings. This coefficient measures molecular similarity within the [0,1] range. In line with the convergence of (α,ϵ), the TC also starts to converge in the 50^{th} generation at a rather small value of $\sim$0.33, reflecting the diversity of the TMC populations in the last generations. This diversity can also be appreciated in the structures of the TMC hits shown in Figure 13, which, along the Pareto front, maximize either one or both target properties.

The evolution of the hits with the GFN2-xTB fitness was also assessed by recomputing the target properties at a higher level of theory. From generations 1, 10, 30, 50, and 100, we extracted all TMCs for which the calculation of α and ϵ with xTB was successful. For these TMCs, 99% of the total, we computed α and ϵ at the DFT(PBE/def2SVP) and DFT(PBE0/def2TZVP) levels, respectively. The results were plotted against the 60K tmQMg dataset, for which these properties were also available at the same DFT level (Figure S9). This plot showed that 1) there was clear progress towards the DFT tmQMg Pareto front, though less stable than with xTB (Figure 11) and in line with the significant differences between these two methods, and 2) there was a gap between the TMC hits and the DFT tmQMg Pareto front, in line with the different sizes of the associated fragment pools; *i.e.*: 1 *vs.* 30, for the metal center, and 50 *vs.* 30K for the ligands. We hypothesized that the average deviation between DFT and xTB would increase by approaching the Pareto front but the opposite trend was seen, likely due to the TMCs being constrained into a smaller (α,ϵ) region. These results also suggested that an xTB-guided MOGA followed by a DFT verification of selected solutions may constitute a robust computational protocol.

PL-MOGA benchmark

Being able to guide the MOGA search towards one specific region of the Pareto front can be of high interest as this allows biasing the TMCs of the final population to emphasize certain properties more than others (Figure 3). This can be useful in applications where one property is particularly important and thus its optimization should be prioritized. One way of achieving this is to introduce a masking function that sets the fitness vector of any TMC to zero if one of its components is lower than a specified threshold. This causes those TMCs to be strongly disfavored during parent and survivor selection, thereby ensuring that the final population is pushed towards fitness values higher than the threshold. Previously, this has been done with static thresholds that stay constant during the whole MOGA run.[68,69] However, this requires setting an appropriate threshold *a priori* without knowledge of the ranges of the fitness values and is therefore almost exclusively useful for constraint handling.

We hereby propose a novel dynamic masking procedure that keeps updating the threshold values based on the median over the current population. In each generation, the function loops over a list of selected targets, considering their values for masking. For each of these targets, the population median is calculated and multiplied by a target-specific scaling factor to obtain a threshold. The target values of each individual are then compared to these thresholds and if any of them is smaller, their fitness target vector is set to zero. This ensures an additional, continuous selection pressure and pushes the population towards a specific region of the Pareto front. The scaling factors can be chosen individually for each target and, depending on the choice of values, the position (*i.e.* aim) as well as the width (*i.e.* scope) of the region explored can be tuned smoothly. We experimented with different values and found that the scaling factors should be within the continuous $[0, 1]$ range, where zero corresponds to no masking applied. Scaling values larger than 1 yielded high thresholds with which large portions of the population were mapped to zero fitness, significantly hindering evolution.

Algorithm 1 PL-MOGA ZEROMASK function

function ZEROMASK(x, X, T, S)
  for t in T **do**
    $m_t \leftarrow$ population X_t median
    if $x_t < s_t \cdot m_t$ **then**
      return 0
    end if
  end for
  return x
end function

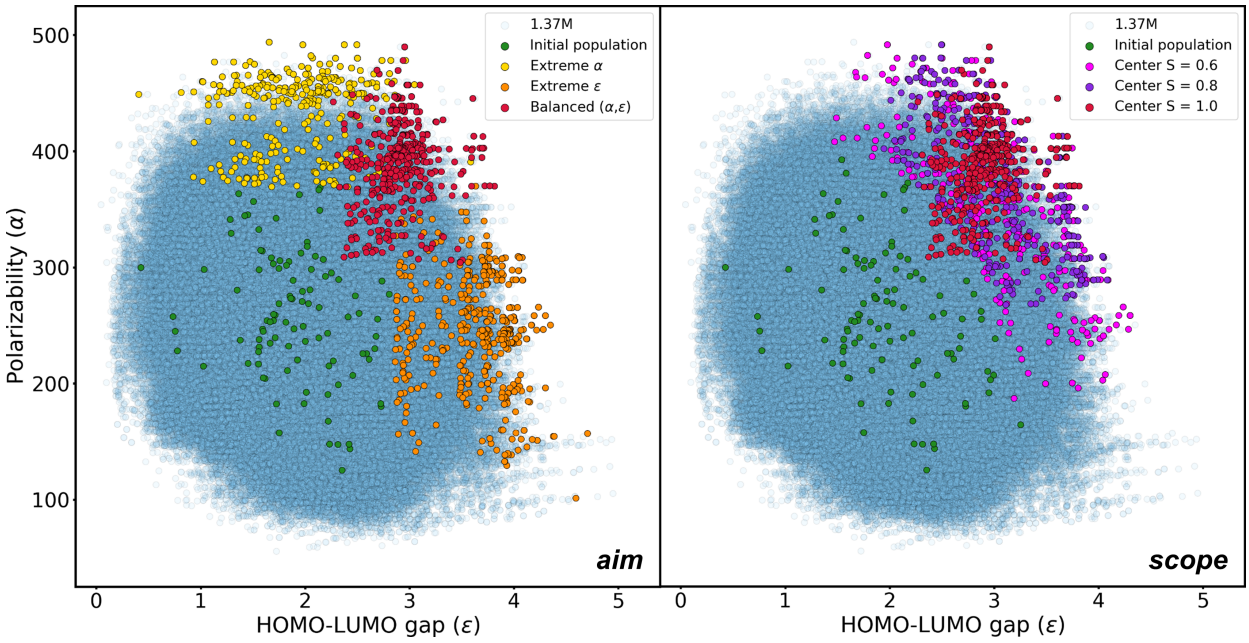

Figure 14: Pareto-Lighthouse MOGA in the 1.37M space. In the aimed optimizations (left-hand side), the (S_α,S_ϵ) scaling factors were (1,0), (0,1), and (1,1) for the extreme α, extreme ϵ, and α,ϵ-balanced runs, respectively. In the scope optimizations (right-hand side), all runs were center-aimed with a widening scope set by the $S = S_\alpha = S_\epsilon$ values shown in the legend. Only the initial random population (the same in both cases) and the final populations are shown. The α and ϵ units are Bohr^3 and eV, respectively.

We called this MOGA *Pareto-Lighthouse* (PL-MOGA), in analogy to a seacoast lighthouse in which the aim and scope of the beam can be modulated, and we implemented it by developing the ZEROMASK function. Algorithm 1 shows the pseudocode of this function, where x denotes the fitness vector of any query TMC, X denotes the list of fitness vectors of the current population, T denotes the list of t indices of the selected targets, and S denotes the list of corresponding factors s_t used to scale the population median of each target (m_t). ZEROMASK returns the fitness either unchanged or transformed to zero, assuming that this is its lowest possible value for any target. This function is only valid when $dim(T) = dim(S)$; *i.e.* one scaling factor must be provided for each selected target.

Figure 14 shows the application of the PL-MOGA algorithm to the optimization of the α and ϵ properties within the 1.37M space. In a first run, with (S_α,S_ϵ) = (1,0), the algorithm explored the Pareto front region in which α and ϵ are maximized and minimized, respectively, whereas the opposite region was explored in a subsequent run after permuting the scaling factors to (S_α,S_ϵ) = (0,1). The center region of the Pareto front, in which both target properties are maximized in a balanced manner, was also explored using (S_α,S_ϵ) = (1,1). Additional "in-between" explorations can be made with $0 < S_\alpha$, $S_\epsilon < 1$, and either $S_\alpha < S_\epsilon$ or $S_\alpha > S_\epsilon$, depending on whether the region of interest is α- or ϵ-biased, respectively. Further, the scope of the calculation can also be fine-tuned; *e.g.*, Figure 14 illustrates how a MOGA optimization aiming at the center of the Pareto front can be gradually widened by decreasing the value of S from 1.0 to 0.8 and 0.6, with $S = S_\alpha = S_\epsilon$. A conventional, *i.e.* unmasked, MOGA optimization can also be easily set with $S = S_\alpha = S_\epsilon = 0$. The PL-MOGA thus allows for exploring the Pareto front in a continuous manner, by controlling both the aim and the scope of the optimization with a simple an intuitive choice of scaling factors.

Billion-scale multiobjective optimization

After benchmarking the PL-MOGA algorithm in the 1.37M space, we scaled up the multiobjective (α,ϵ) optimization task to implicit chemical spaces containing billions of TMCs. Two different 252-ligand pools, both monodentate, were defined: the extended and the random. The extended pool was made by adding 202 ligands of decreasing popularity to the 50-ligand pool underlying the 1.37M space. Popularity (*i.e.* number of occurrences in the tmQMg dataset for Pd complexes) was decreased with the aim of exploring more unusual TMCs. The random pool was made with a random selection of 252 ligands from the tmQMg-L dataset. We used the square planar Pd(II) scaffold, which, for each of these two pools, yields 1,008,189,504 unique TMCs (Equation 2). The same charge constraints of the 1.37M space were applied: {0, -1} for the ligands, in a 1:1 ratio, and {-1, 0, +1} for the resulting TMCs. Regarding ligand size, we kept the same limit in the extended pool ($\leq$15 heavy atoms), whereas, for the random, any ligand size was allowed to maximize the scope of the multiobjective optimization.

The spaces resulting from the extended and random ligand pools were explored by evolving a total of 195,000 TMCs over 150 generations of 1,300 TMCs each. At this scale, both target properties were converged (Figure S10), with the ligand use frequency histograms showing that all 252 ligands in the two pools were used from $\sim$1K to 67K times (Figure S11), after exploring 0.02% of the space. As in the 1.37M space benchmark, these histograms could be used to follow and interpret the evolution of the TMC hits. The average Tanimoto coefficients converged at $\sim$0.2 (Figure S12), showing high chemical diversity in the last generation of TMC hits, which contained 108 (extended pool) and 143 (random pool) unique ligands. This value is significantly smaller than that converged in the 1.37M space ($\sim$0.33 in Figure 13), in line with the larger size of the ligand pools and the associated chemical spaces. The PL-MOGA was also used to show that the 1,300 hits from the last generation could be chosen to be either scattered over the entire Pareto front (*i.e.* no masking) or concentrated at its center with the $(S_\alpha, S_\epsilon) = (1, 1)$ mask.

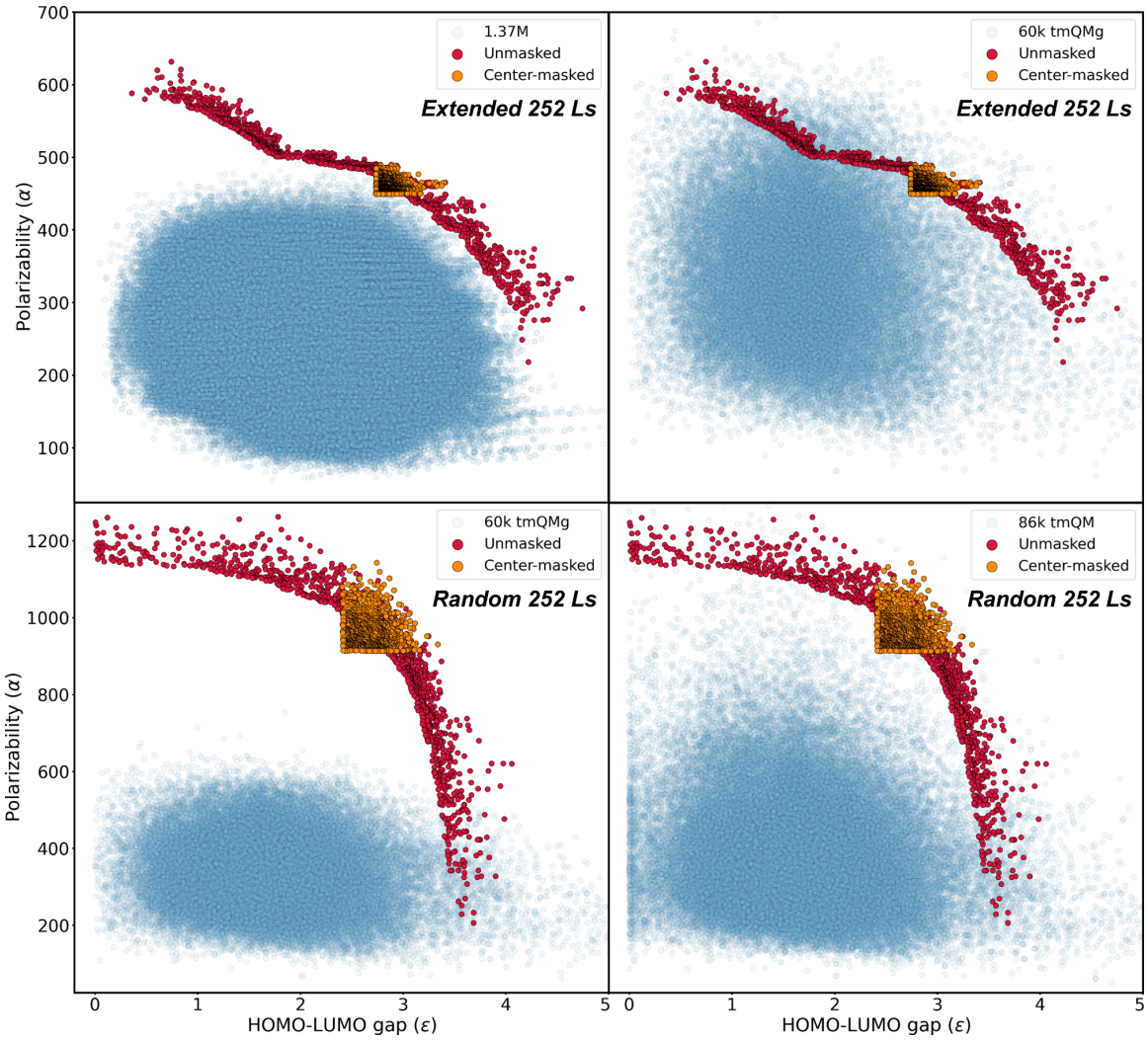

Figure 15: Multiobjective (α,ϵ) optimization in the billion chemical spaces with the PL-MOGA algorithm. The last generation of hits is plotted for both the extended and random 252-ligand pools over the 1.37M, tmQMg, and tmQM spaces. The center-masked optimization was done with $(S_\alpha, S_\epsilon) = (1, 1)$. The α and ϵ units are Bohr^3 and eV, respectively.

Figure 15 shows the (α, ϵ) coordinates of the 1,300 TMC hits evolved from the extended ligand pool. Relative to the 1.37M space, the unmasked PL-MOGA leveraged the 202 ligands added to the original 50-ligand set, yielding a new Pareto front well ahead in the (α,ϵ) map. The TMC hits pushed the upper limit of α from ~450 to 600 Bohr^3, despite the ligand size limit. Further, the ~600 Bohr^3 limit coincides with that of the whole tmQMg space, which contains larger TMCs up to a maximum size of 85 atoms. These results can be ascribed to the selection of cyclic ligands containing heavier elements like sulfur, as shown in Figure

16 for a random selection of TMC hits. The hits covered most of the tmQMg Pareto front, using only the 252 ligands available from the pool, which was much smaller than the 30K tmQMg-L dataset underlying this chemical space. At the $\epsilon > 3$ eV extreme, the algorithm picked strong field ligands (*e.g.* $CF_3(CF_2)_3^-$), adding also larger rings and heavier elements in the central region of the Pareto front, where both α and ϵ were maximized jointly to ∼2.5 eV and ∼450 Bohr^3, respectively. This same central region was populated exclusively by using the center mask – the small number of new dominating points generated in this calculation suggested that a physical limit was reached with this ligand pool, which, in the unmasked run, was also incapable of reaching the region at ∼2 eV / 550 Bohr^3. Two repetitions of this calculation from different random initial populations yielded the same observation, producing hits at very similar (α, ϵ) coordinates and thus also reflecting the robustness of the algorithm (Figure S13).

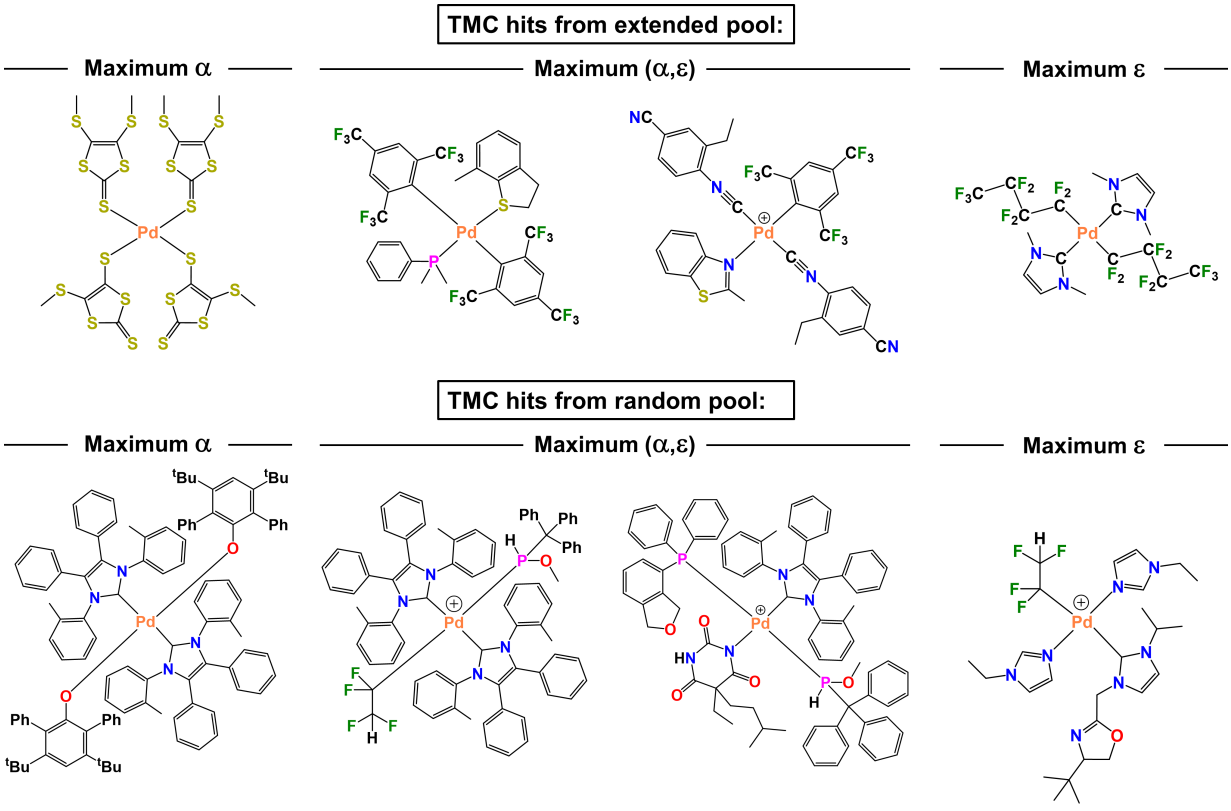

Figure 16: Random samples of the TMC hits evolved within the billion spaces. Long lines were only used to simplify the 2D drawings and they do not represent unusually long bonds in any case. The structures of the largest TMCs were verified at the DFT(PBE/def2SVP) level (SI).

With the random pool, we first took tmQMg as reference since this chemical space is based on the same ligand dataset (*i.e.* tmQMg-L) from which the 252 ligands were extracted. The plot in Figure 15 shows how the evolved TMC hits redefined most of the Pareto front, with the exception of the ϵ extreme, for which only a few hits reached $\epsilon \geq 4$ eV. This observation may signal that further sampling is needed to cover this region by, for example, repeating the optimization with a different set of random ligands. Regarding α, the hits opened a wide gap, which at $\epsilon \sim 2.5$ eV spans $\sim$400 Bohr^3. This dramatic gap is mostly due to molecular size, which was limited to 85 atoms in tmQMg, whereas the PL-MOGA could produce hits with more than 200 atoms. We thus considered a second chemical space as a reference: tmQM, the original dataset of this series.[58] tmQM included 86K mononuclear TMCs from the CSD database, selecting the highest quality structures with a charge in the [-1, 0, +1] range, without imposing any size filter (the largest TMC in tmQM has 569 atoms).

Figure 15 shows that the population of hits evolved from the random pool with the unmasked PL-MOGA largely overlaps with the tmQM Pareto front, augmenting the density of TMCs in this region significantly. At the $\alpha \sim 1200$ Bohr extreme, the algorithm optimized very large TMCs like the one shown in Figure 16. We also observed that in several TMCs the ligands were fused through bond rearrangements yielding systems that can be chemically valid, converging an optimized geometry at the xTB level, but probably very unstable, since, in this maximum α region, ϵ is minimized to $\sim 0 - 0.5$ eV. Besides this region, the MOGA found hundreds of stable TMCs maximizing either ϵ alone or both ϵ and α. The hits shown in Figure 16 for these two regions of the Pareto front, which were all verified at the DFT(PBE/def2SVP) level (Figure S14), show how the algorithm managed to leverage the random 252-ligand pool to select and combine both common (*e.g.* phosphines and carbenes) and unusual (*e.g.* alkoxides and alkyls) ligands for the Pd(II) scaffold. Further, all 1,300 TMC hits were new relative to the tmQM dataset, which was also the case when the center mask was applied. Hence, the PL-MOGA was capable of generating novel TMC spaces in a directional multiobjective optimization framework.

Conclusions

The present work showed how NBO and graph theories enabled the robust definition of TMC ligands in terms of their charge and metal coordination mode. With this approach, we curated the 30K tmQMg-L ligand dataset, which combines synthesizability with a wide range of molecular sizes and chemical environments, in a format enabling the automated exploration of the vast TMC chemical space. The information provided by tmQMg-L was leveraged in the generation of a TMC space based on the square planar palladium(II) scaffold. Using only 50 of the 30K ligands available and doing the full combinatorial explosion, we obtained a chemical space of 1.37M unique TMCs, in which two quantum properties, *i.e.* the polarizability (α) and the HOMO-LUMO gap (ϵ), appeared uncorrelated over wide ranges ($\sim 50 - 475$ Bohr^3 and $0.15 - 4.15$ eV), forming a Pareto front.

Using this chemical space as a benchmark, we developed a MOGA for the multiobjective optimization of TMCs. The evolution of the hits was implemented through full-ligand genetic operations based on the geometry and isomerism of the square planar scaffold, in a way that can be easily extended to other metal coordination geometries. In the 1.37M space, the MOGA located 130 TMC hits over the (α, ϵ) Pareto front with high chemical diversity and in an explainable manner. This approach was extended with the PL-MOGA algorithm, which allowed for defining the aim and scope of the optimizer over the Pareto front through the intuitive selection of scaling factors. When the optimization task was scaled up to implicit spaces containing billions of TMCs, the PL-MOGA located thousands of hits that were novel relative to extensive datasets extracted from the CSD, showing its potential for the discovery of TMCs within unexplored regions of the chemical space.

Together, the tmQMg-L dataset and the PL-MOGA algorithm constitute a robust generative method for TMCs optimizing multiple properties based on the choice of multiple ligands. We envision that this method can be the basis of an evolutionary learning strategy in which, first, multiple ligands are selected to optimize properties, that, in a subsequent step, are further refined with GAs acting at the atomic level on individual ligands.

Supporting information

The Supporting Information provides further details about the tmQMg-L dataset, the 1.37M chemical space, the PL-MOGA algorithm, the estimation of the chemical diversity with average Tanimoto coefficients, the DFT benchmark, repetitions from different random initial populations, additional information on the exploration of the implicit billion spaces, and general computational and chemoinformatics details.

Data and code

All data and code are openly available. The tmQMg-L dataset can be accessed at the URL: *https://github.com/hkneiding/tmQMg-L* whereas the PL-MOGA code is available from the URL: *https://github.com/hkneiding/PL-MOGA*, which also provides the DFT geometries of selected TMC hits.

Author contributions

HK was the main developer of the tmQMg-L dataset, the 1.37M space, and the MOGA, including the Pareto-Lighthouse algorithm. HK also derived the combinatorics of the square planar TMC space and developed the concept of a generative model based on a MOGA acting at the multiligand whole-complex level. AN and DB developed the concept of extracting the ligand charges from the natural Lewis structures. All authors made substantial contributions to the conception and design of the work. DB was the main contributor to the writing and revision of the manuscript, as well as to the definition, supervision, and funding of the research project.

Conflicts of interest

There are no conflicts of interest to declare.

Acknowledgements

HK acknowledges the support from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No. 945371. This article reflects only the author's view and the REA is not responsible for any use that may be made of the information it contains. AN acknowledges the support from the Research Council of Norway (RCN) through its FRIPRO program (CO2pCat project; number 314321). DB also acknowledges the support from the RCN FRIPRO program (catLEGOS project; number 325003). AN and DB acknowledge the RCN for its support through the Centers of Excellence program (Hylleraas Centre; project number 262695) and the Norwegian Supercomputing Program (NOTUR; project number NN4654K). We thank Magnus Strandgaard for reviewing a preliminary version of this manuscript.

References

(1) Mjos, K. D.; Orvig, C. Metallodrugs in Medicinal Inorganic Chemistry. *Chem. Rev.* **2014**, *114*, 4540–4563.

(2) Prier, C. K.; Rankic, D. A.; MacMillan, D. W. C. Visible Light Photoredox Catalysis with Transition Metal Complexes: Applications in Organic Synthesis. *Chem. Rev.* **2013**, *113*, 5322–5363.

(3) Kalyanasundaram, K.; Gratzel, M. Applications of functionalized transition metal complexes in photonic and optoelectronic devices. *Coord. Chem. Rev.* **1998**, *177*, 347–414.

(4) Yoon, T. P.; Ischay, M. A.; Du, J. N. Visible light photocatalysis as a greener approach to photochemical synthesis. *Nature Chem.* **2010**, *2*, 527–532.

(5) Furukawa, H.; Cordova, K. E.; O'Keeffe, M.; Yaghi, O. M. The Chemistry and Applications of Metal-Organic Frameworks. *Science* **2013**, *341*, 974.

(6) Sperger, T.; Sanhueza, I. A.; Kalvet, I.; Schoenebeck, F. Computational Studies of Synthetically Relevant Homogeneous Organometallic Catalysis Involving Ni, Pd, Ir, and Rh: An Overview of Commonly Employed DFT Methods and Mechanistic Insights. *Chem. Rev.* **2015**, *115*, 9532–9586.

(7) Balcells, D.; Nova, A. Designing Pd and Ni Catalysts for Cross-Coupling Reactions by Minimizing Off-Cycle Species. *ACS Catal.* **2018**, *8*, 3499–3515.

(8) Foscato, M.; Jensen, V. R. Automated in Silico Design of Homogeneous Catalysts. *ACS Catal.* **2020**, *10*, 2354–2377.

(9) Robbins, D. W.; Hartwig, J. F. A Simple, Multidimensional Approach to High-Throughput Discovery of Catalytic Reactions. *Science* **2011**, *333*, 1423–1427.

(10) Nandy, A.; Duan, C. R.; Taylor, M. G.; Liu, F.; Steeves, A. H.; Kulik, H. J. Computational Discovery of Transition-metal Complexes: From High-throughput Screening to Machine Learning. *Chem. Rev.* **2021**, *121*, 9927–10000.

(11) Huang, B.; von Lilienfeld, O. A. Ab Initio Machine Learning in Chemical Compound Space. *Chem. Rev.* **2021**, *121*, 10001–10036.

(12) Freeze, J. G.; Kelly, H. R.; Batista, V. S. Search for Catalysts by Inverse Design: Artificial Intelligence, Mountain Climbers, and Alchemists. *Chem. Rev.* **2019**, *119*, 6595–6612.

(13) Kitchin, J. R. Machine learning in catalysis. *Nat. Catal.* **2018**, *1*, 230–232.

(14) Gomes, G. D.; Pollice, R.; Aspuru-Guzik, A. Navigating through the Maze of Homogeneous Catalyst Design with Machine Learning. *Trends Chem.* **2021**, *3*, 96–110.

(15) Pablo-Garcia, S.; Morandi, S.; Vargas-Hernandez, R. A.; Jorner, K.; Ivkovic, Z.; Lopez, N.; Aspuru-Guzik, A. Fast evaluation of the adsorption energy of organic molecules on metals via graph neural networks. *Nat. Comput. Sci.* **2023**, *3*, 433–442.

(16) Friederich, P.; Gomes, G. D.; De Bin, R.; Aspuru-Guzik, A.; Balcells, D. Machine learning dihydrogen activation in the chemical space surrounding Vaska's complex. *Chem. Sci.* **2020**, *11*, 4584–4601.

(17) Nandy, A.; Duan, C. R.; Goffinet, C.; Kulik, H. J. New Strategies for Direct Methane-to-Methanol Conversion from Active Learning Exploration of 16 Million Catalysts. *JACS Au* **2022**, *2*, 1200–1213.

(18) Cordova, M.; Wodrich, M. D.; Meyer, B.; Sawatlon, B.; Corminboeuf, C. Data-Driven Advancement of Homogeneous Nickel Catalyst Activity for Aryl Ether Cleavage. *ACS Catal.* **2020**, *10*, 7021–7031.

(19) Jorner, K.; Tomberg, A.; Bauer, C.; Skold, C.; Norrby, P. O. Organic reactivity from mechanism to machine learning. *Nat. Rev. Chem.* **2021**, *5*, 240–255.

(20) Vamathevan, J.; Clark, D.; Czodrowski, P.; Dunham, I.; Ferran, E.; Lee, B., G ans Li; Madabhushi, A.; Shah, P.; ; Spitzer, M. Applications of machine learning in drug discovery and development. *Nat. Rev. Drug Discov.* **2019**, *18*, 463–477.

(21) Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine learning for molecular and materials science. *Nature* **2018547-555**, *559*, 547–555.

(22) Goldberg, D. E. *Genetic Algorithms in Search, Optimization, and Machine Learning*; Addison-Wesley, 1986.

(23) De Jong, K. A. *Evolutionary Computation – A Unified Appraoch*; The MIT Press, 2006.

(24) Winter, R.; Montanari, F.; Steffen, A.; Briem, H.; Noe, F.; Clevert, D. A. Efficient multi-objective molecular optimization in a continuous latent space. *Chem. Sci.* **2019**, *10*, 8016–8024.

(25) Le, T. C.; Winkler, D. A. Discovery and Optimization of Materials Using Evolutionary Approaches. *Chem. Rev.* **2016**, *116*, 6107–6132.

(26) Anstine, D. M.; Isayev, O. Generative Models as an Emerging Paradigm in the Chemical Sciences. *J. Am. Chem. Soc.* **2023**, *145*, 8736–8750.

(27) Sanchez-Lengeling, B.; Aspuru-Guzik, A. Inverse molecular design using machine learning: Generative models for matter engineering. *Science* **2018**, *361*, 360–365.

(28) Jensen, J. H. A graph-based genetic algorithm and generative model/Monte Carlo tree search for the exploration of chemical space. *Chem. Sci.* **2019**, *10*, 3567–3572.

(29) Nigam, A.; Pollice, A.; Aspuru-Guzik, A. Parallel tempered genetic algorithm guided by deep neural networks for inverse molecular design. *Digital Discovery* **2022**, *1*, 390–404.

(30) Janet, J. P.; Chan, L.; Kulik, H. J. Accelerating Chemical Discovery with Machine Learning: Simulated Evolution of Spin Crossover Complexes with an Artificial Neural Network. *J. Phys. Chem. Lett.* **2018**, *9*, 1064–1071.

(31) Jennings, P. C.; Lysgaard, S.; Hummelshoj, J. S.; Vegge, T.; Bligaard, T. Genetic algorithms for computational materials discovery accelerated by machine learning. *Npj Comput. Mater.* **2019**, *5*, 46.

(32) Gallarati, S.; Gerwen, P. v.; Schoepfer, A. A.; Laplaza, R.; Corminboeuf, C. Genetic Algorithms for the Discovery of Homogeneous Catalysts. *CHIMIA* **2023**, *77*, 39.

(33) Fey, N.; Orpen, A. G.; Harvey, J. N. Building ligand knowledge bases for organometallic chemistry: Computational description of phosphorus(III)-donor ligands and the metal-phosphorus bond. *Coord. Chem. Rev.* **2009**, *253*, 704–722.

(34) Fey, N.; Haddow, M. F.; Harvey, J. N.; McMullin, C. L.; Orpen, A. G. A ligand knowledge base for carbenes (LKB-C): maps of ligand space. *Dalton Trans.* **2009**, 8183–8196.

(35) Gugler, S.; Janet, J. P.; Kulik, H. J. Enumeration of de novo inorganic complexes for chemical discovery and machine learning. *Mol. Syst. Des. Eng.* **2020**, *5*, 139–152.

(36) Gensch, T.; Gomes, G. D.; Friederich, P.; Peters, E.; Gaudin, T.; Pollice, R.; Jorner, K.; Nigam, A.; Lindner-D'Addario, M.; Sigman, M. S. A Comprehensive Discovery Platform for Organophosphorus Ligands for Catalysis. *J. Am. Chem. Soc.* **2022**, *144*, 1205–1217.

(37) Ioannidis, E. I.; Gani, T. Z. H.; Kulik, H. J. molSimplify: A toolkit for automating discovery in inorganic chemistry. *J. Comput. Chem.* **2016**, *37*, 2106–2117.

(38) Foscato, M.; Venkatraman, V.; Jensen, V. R. DENOPTIM: Software for Computational de Novo Design of Organic and Inorganic Molecules. *J. Chem. Inf. Model.* **2019**, *59*, 4077–4082.

(39) Guan, Y. F.; Ingman, V. M.; Rooks, B. J.; Wheeler, S. E. AARON: An Automated Reaction Optimizer for New Catalysts. *J. Chem. Theory Comput.* **2018**, *14*, 5249–5261.

(40) Sobez, J. G.; Reiher, M. MOLASSEMBLER: Molecular Graph Construction, Modification, and Conformer Generation for Inorganic and Organic Molecules. *J. Chem. Inf. Model.* **2020**, *60*, 3884–3900.

(41) Chen, S.; Nielson, T.; Zalit, E.; Skjelstad, B. B.; Borough, B.; Hirschi, W. J.; Yu, S.; Balcells, D.; Ess, D. H. Automated Construction and Optimization Combined with Machine Learning to Generate Pt(II) Methane C-H Activation Transition States. *Top. Catal.* **2022**, *65*, 312–324.

(42) Kneiding, H.; Lukin, R.; Lang, L.; Reine, S.; Pedersen, T. B.; De Bin, R.; Balcells, D. Deep learning metal complex properties with natural quantum graphs. *Digital Discovery* **2023**, *2*, 618–633.

(43) Groom, C. R.; Bruno, I. J.; Lightfoot, M. P.; Ward, S. C. The Cambridge Structural Database. *Acta Cryst. B* **2016**, *B72*, 171–179.

(44) Duan, C.; Ladera, A. J.; Liu, J.; Taylor, M. G.; Ariyarathna, I. R.; Kulik, H. J. Exploiting Ligand Additivity for Transferable Machine Learning of Multireference Character across Known Transition Metal Complex Ligands. *J. Chem. Theory Comput.* **2022**, *18*, 4836–4845.

(45) Vela, S.; Laplaza, R.; Cho, Y. R.; Corminboeuf, C. cell2mol: encoding chemistry to interpret crystallographic data. *Npj Comput. Mater.* **2022**, *8*, 188.

(46) Matsuoka, W.; Harabuchi, Y.; Maeda, S. Virtual Ligand-Assisted Screening Strategy to Discover Enabling Ligands for Transition Metal Catalysis. *ACS Catal.* **2022**, *12*, 3752–3766.

(47) Gao, W. H.; Coley, C. W. The Synthesizability of Molecules Proposed by Generative Models. *J. Chem. Inf. Model.* **2020**, *60*, 5714–5723.

(48) Foscato, M.; Occhipinti, G.; Venkatraman, V.; Alsberg, B. K.; Jensen, V. R. Automated Design of Realistic Organometallic Molecules from Fragments. *J. Chem. Inf. Model.* **2014**, 767–780.

(49) Chu, Y. H.; Heyndrickx, W.; Occhipinti, G.; Jensen, V. R.; Alsberg, B. K. An Evolutionary Algorithm for de Novo Optimization of Functional Transition Metal Compounds. *J. Am. Chem. Soc.* **2012**, *134*, 8885–8895.

(50) Durrant, M. C. The Use of Quantum Molecular Calculations to Guide a Genetic Algorithm: A Way to Search for New Chemistry. *Chem. Eur. J.* **2007**, *13*, 3406–3413.

(51) Verhellen, J. Graph-based molecular Pareto optimisation. *Chem. Sci.* **2022**, *13*, 7526–7535.

(52) Hase, F.; Roch, L. M.; Aspuru-Guzik, A. Chimera: enabling hierarchy based multi-objective optimization for self-driving laboratories. *Chem. Sci.* **2018**, *9*, 7642–7655.

(53) Nigam, A.; Pollice, R.; Krenn, M.; Gomes, G. D.; Aspuru-Guzik, A. Beyond generative models: superfast traversal, optimization, novelty, exploration and discovery (STONED) algorithm for molecules using SELFIES. *Chem. Sci.* **2021**, *12*, 7079–7090.

(54) Laplaza, R.; Gallarati, S.; Corminboeuf, C. Genetic Optimization of Homogeneous Catalysts. *Chemistry–Methods* **2022**, *2*, e202100107.

(55) Janet, J. P.; Ramesh, S.; Duan, C.; Kulik, H. J. Accurate Multiobjective Design in a Space of Millions of Transition Metal Complexes with Neural-Network-Driven Efficient Global Optimization. *ACS Cent. Sci.* **2020**, *6*, 513–524.

(56) Sowndarya, S. V. S.; Law, J. N.; Tripp, C. E.; Duplyakin, D.; Skordilis, E.; Biagioni, D.; Paton, R. S.; St John, P. C. Multi-objective goal-directed optimization of de novo stable organic radicals for aqueous redox flow batteries. *Nat. Mach. Intell.* **2022**, *4*, 720–730.

(57) Seumer, J.; Hansen, J. K. S.; Nielsen, M. B.; Jensen, J. H. Computational Evolution Of New Catalysts For The Morita-Baylis-Hillman Reaction. *Angew. Chem. Int. Ed.* **2023**, e202218565.

(58) Balcells, D.; Skjelstad, B. B. tmQM Dataset—Quantum Geometries and Properties of 86k Transition Metal Complexes. *J. Chem. Inf. Model.* **2020**, *60*, 6135–6146.

(59) Chen, S.; Meyer, Z.; Jensen, B.; Kraus, A.; Lambert, A.; Ess, D. H. ReaLigands: A Ligand Library Cultivated from Experiment and Intended for Molecular Computational Catalyst Design. *ChemRxiv* **2023**, *preprint*, 10.26434/chemrxiv–2023–ngqd8.

(60) Sauer, W. H.; Schwarz, M. K. Molecular shape diversity of combinatorial libraries: a prerequisite for broad bioactivity. *J. Chem. Inf. Model.* **2003**, *43*, 987–1003.

(61) Shrake, A.; Rupley, J. A. Environment and exposure to solvent of protein atoms. Lysozyme and insulin. *J. Mol. Biol.* **1973**, *79*, 351–371.

(62) Eisenhaber, F.; Lijnzaad, P.; Argos, P.; Sander, C.; Scharf, M. The double cubic lattice method: Efficient approaches to numerical integration of surface area and volume and to dot surface contouring of molecular assemblies. *J. Comput. Chem.* **1995**, *16*, 273–284.

(63) Falivene, L.; Credendino, R.; Poater, A.; Petta, A.; Serra, L.; Oliva, R.; Scarano, V.; Cavallo, L. SambVca 2. A web tool for analyzing catalytic pockets with topographic steric maps. *Organometallics* **2016**, *35*, 2286–2293.

(64) Tolman, C. A. Phosphorus ligand exchange equilibriums on zerovalent nickel. Dominant role for steric effects. *J. Am. Chem. Soc.* **1970**, *92*, 2956–2965.

(65) Bilbrey, J. A.; Kazez, A. H.; Locklin, J.; Allen, W. D. Exact ligand cone angles. *J. Comput. Chem.* **2013**, *34*, 1189–1197.

(66) Bilbrey, J. A.; Kazez, A. H.; Locklin, J.; Allen, W. D. Exact Ligand Solid Angles. *J. Chem. Theory Comput.* **2013**, *9*, 5734–5744.

(67) Guzei, I. A.; Wendt, M. An improved method for the computation of ligand steric effects based on solid angles. *Dalton Trans.* **2006**, 3991–3999.

(68) Hoffmeister, F.; Sprave, J. Problem-Independent Handling of Constraints by Use of Metric Penalty Functions. *Evolutionary Programming* **1996**, *870*.

(69) Devi, R. V.; Sathya, S. S.; Coumar, M. S. Multi-objective genetic algorithm for De novo drug design (MoGADdrug). *Current Computer-Aided Drug Design* **2021**, *17*, 445–457.