CATALYSIS

Accelerated dinuclear palladium catalyst identification through unsupervised machine learning

Julian A. Hueffel[1]†, Theresa Sperger[1]†, Ignacio Funes-Ardoiz[1], Jas S. Ward[2], Kari Rissanen[2], Franziska Schoenebeck[1]*

Although machine learning bears enormous potential to accelerate developments in homogeneous catalysis, the frequent need for extensive experimental data can be a bottleneck for implementation. Here, we report an unsupervised machine learning workflow that uses only five experimental data points. It makes use of generalized parameter databases that are complemented with problem-specific in silico data acquisition and clustering. We showcase the power of this strategy for the challenging problem of speciation of palladium (Pd) catalysts, for which a mechanistic rationale is currently lacking. From a total space of 348 ligands, the algorithm predicted, and we experimentally verified, a number of phosphine ligands (including previously never synthesized ones) that give dinuclear $Pd^{(I)}$ complexes over the more common $Pd^{(0)}$ and $Pd^{(II)}$ species.

The speciation of homogeneous metal catalysts is a key determinant of reactivity, efficiency, and selectivity. Yet, the factors that dictate nuclearity (e.g., monomer versus dimer), favored oxidation state, and ligation state of a catalyst are all too frequently barely understood (*1*). For example, in the context of the widely employed Pd-catalyzed cross-coupling, the oxidative addition of $Pd^{(0)}[P(t\text{-}Bu)_3]_2$ to an aryl bromide generates a $Pd^{(I)}$ dimer in situ (*2*), and the same $Pd^{(I)}$ dimer is converted to a Pd trimer when $P(t\text{-}Bu)_3$ is displaced by Ph_2PH (*3*). This speciation challenge is further aggravated for nonprecious metal species (e.g., Fe, Co, Cu, or Ni catalysts), for which subtle ligand differences may affect the favored spin state in addition to oxidation state and nuclearity (*4*). Clearly, the nature of the coordinating ligands critically influences the speciation. However, the underlying origin of each ligand's impact is rarely understood or predictable (Fig. 1).

Without insight into the correlation between ligand and catalyst speciation, the development of new catalysts heavily relies on trial and error or high-throughput screening efforts. Whereas the former approach tends to be biased by intuition, the latter is dependent on availability or ease of accessibility of vast ligand libraries. Any departure into unknown ligand space is then confronted with the challenge of choice from the vast structural possibilities. There has therefore been a long-standing interest in predicting the likely impact of a given ligand on structure and reactivity of organometallic complexes using parameterization and as such offering qualitative guidance in the available ligand space. Early examples in this context include the widely used Tolman cone angle (*5*) as a metric for a ligand's steric impact, or CO stretching frequency as a measure of a ligand's electronic influence (*6*, *7*). In recent years, developments have attempted to build more comprehensive representations through parameterization of phosphine ligands with a set of descriptors (*8*, *9*). For example, Fey, Harvey, Orpen and co-workers developed a set of "ligand knowledge bases" (LKBs) (*10*). For monodentate phosphines, the "LKB-P" (see supplementary materials for more details) (*11*, *12*) includes 348 phosphine ligands characterized in silico with 28 different descriptors, ranging from certain ligand-specific data, such as proton affinity or highest occupied/lowest unoccupied molecular orbital (HOMO/LUMO) energies, to calculated data that describe the ligand's interaction in model complexes (coordination to Au, Pt, Pd, B). The collective dataset was subsequently subjected to principal component analysis (*13*) to lower its dimensionality and to provide two-dimensional (2D) maps (i.e., LKB-P in Fig. 1C), wherein ligands with similar general properties reside in a similar area. However, although this representation offers many general insights as well as details regarding electronic and steric properties of a given ligand, the ligand–speciation relationship cannot be derived, as exemplified below. There is hence a need for a fundamentally distinct approach.

Under the assumption that structure and reactivity are intimately connected, and similar structures should therefore share similar reactivity patterns (i.e., as dimer, monomer, in specific coordination motifs, oxidation state, or spin state), we set out to identify a means to predict such correlations. To accomplish this goal, we used machine learning and tested the feasibility of this approach on the $Pd^{(0)}/Pd^{(II)}$ monomer versus $Pd^{(I)}$ dimer speciation challenge.

Although the vast majority of Pd-catalyzed cross-coupling reactions have been ascribed to mononuclear and even-numbered oxidation state catalysts (such as $Pd^{(0)}/Pd^{(II)}$) (*1*, *14*), with certain ligands, dinuclear $Pd^{(I)}$ complexes are formed in situ through comproportionation or oxidation with commonly used additives (*15*). Depending on the precise structure of the $Pd^{(I)}$ dimer, different impacts on the reactivity and efficiency result (*15*). In this context, the dihalide-bridged structural motif $[Pd^{(I)}(\mu\text{-}X)P(t\text{-}Bu)_3]_2$ (Fig. 1B) stands out in terms of its stability and catalytic performance, having been shown to function in situ as an efficient off-cycle precursor to low-coordinate $Pd^{(0)}$ or $Pd^{(II)}$-H species (*15*–*17*). Alternatively, the dimer can also react directly via dinuclear cycles, which are associated with distinct driving forces and practicability, facilitating bond formations that are not readily amenable to traditional $Pd^{(0)}/Pd^{(II)}$ cycles (*18*). These characteristics allow, for example, the a priori predictable control of site selectivity in poly(pseudo)halogenated arenes, even under tolerance of oxygen (*19*).

Although close analogs of $[Pd^{(I)}(\mu\text{-}X)P(t\text{-}Bu)_3]_2$ have been synthesized [i.e., $P(t\text{-}Bu)_2Ph$, $P(1\text{-}Ad)_2(n\text{-}Bu)$, and $P(t\text{-}Bu)_2(i\text{-}Pr)$] (*15*), these new dimers were predominantly developed on a trial and error basis. Indeed, to date, there is little understanding of why certain ligands stabilize $Pd^{(I)}$, whereas others do not. Moreover, it is also unknown why one $Pd^{(I)}$ dimer geometry might be favored over another for a given ligand (*15*). For example, the preferred synthetic approach to $Pd^{(I)}$ dimers consists of comproportionation of $Pd^{(0)}L_2$ with $Pd^{(II)}I_2$, which occurs in the time of mixing at room temperature and is quantitative using the trialkylphosphine ligands mentioned above (*15*). In line with these observations, computational study of the corresponding driving force for comproportionation of $Pd^{(0)}L_2$ with $Pd^{(II)}I_2$ revealed substantial exergonicities of roughly −50 kcal/mol (see Fig. 1C), reinforcing that this process should be highly favored (*20*). However, when considering the trialkylphosphine ligand tricyclohexylphosphine (PCy_3), we calculated the analogous pronounced driving force for comproportionation (i.e., −51.6 kcal/mol, Fig. 1C), yet all experimental attempts to make the corresponding $Pd^{(I)}$ dimer have so far met with failure (*21*). The attempted comproportionation of $Pd^{(0)}(PCy_3)_2$ with $Pd^{(II)}I_2$ instead gives rise to a PCy_3-coordinated $Pd^{(II)}$ dimer—i.e., $[Pd^{(II)}(\mu\text{-}I)(I)PCy_3]_2$—along with precipitation of $Pd^{(0)}$.

This example clearly showcases the complexity of the problem at hand. To accurately predict the favored speciation of catalysts on the basis of mechanistic and quantum mechanical considerations, it is necessary to have precise knowledge of the various potential species in solution that may (or may not) form, their coordination states (with or

[1]Institute of Organic Chemistry, RWTH Aachen University; Landoltweg 1, 52074 Aachen, Germany. [2]Department of Chemistry, University of Jyväskylä; P.O. Box 35, 40014 Jyväskylä, Finland.
*Corresponding author. Email: franziska.schoenebeck@rwth-aachen.de
†These authors contributed equally to this work.

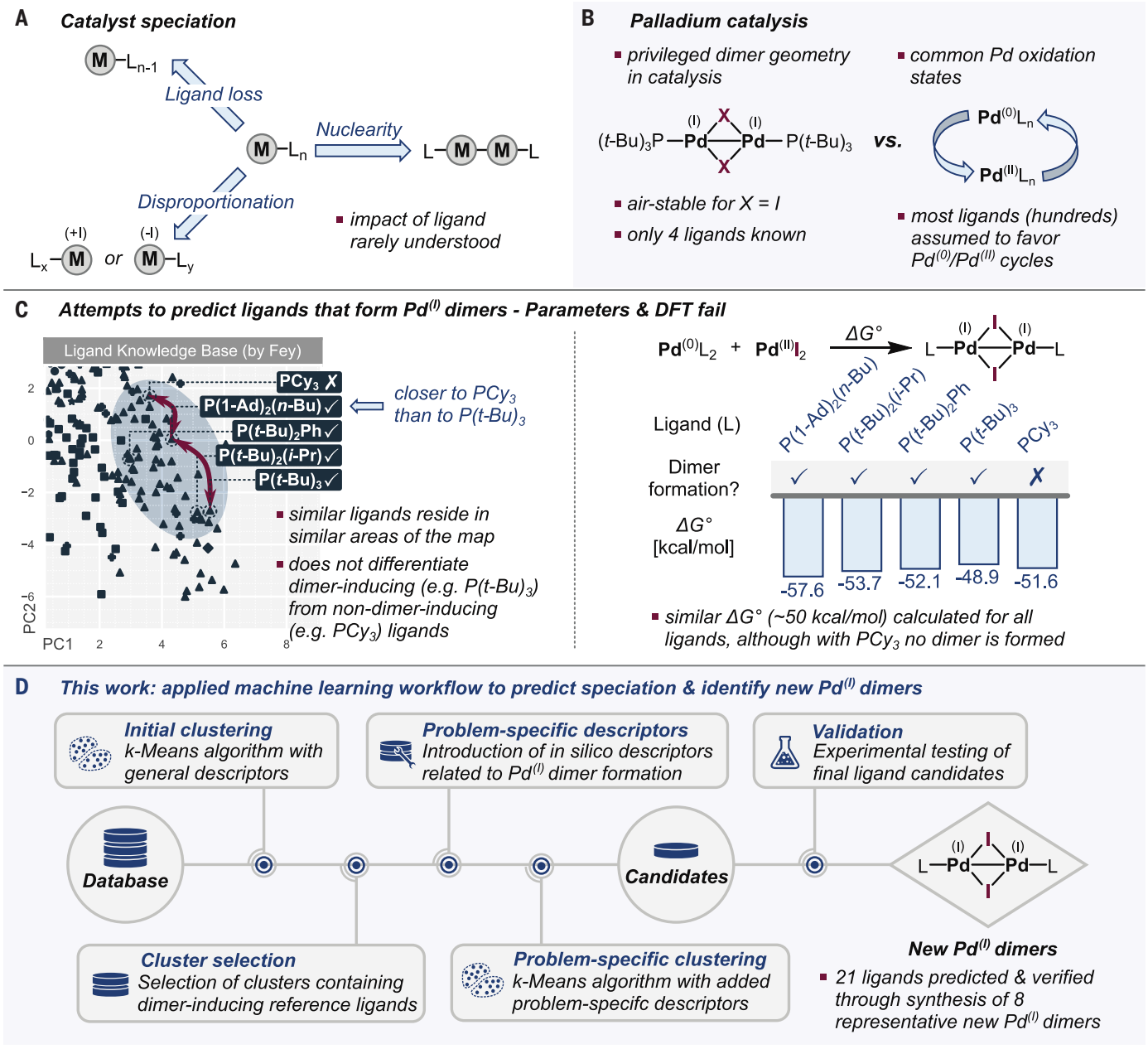

Fig. 1. Context of this work. (**A**) Speciation challenge in catalysis. (**B**) Considered speciation in this work. (**C**) Illustration that current ligand maps or mechanistic data do not allow prediction of dimer speciation. (**D**) Applied machine learning workflow to predict new $Pd^{(I)}$ dimers (this work).

without solvent), spin or charge states, and potential dynamic interconversions. Such information is rarely accessible in full, and it is therefore not surprising that there is to date so little understanding of the factors that dictate catalyst speciation.

Similarly, current ligand databases also offer little insight into which ligands are optimal for dimer formation. When tracking down the currently known $Pd^{(I)}$ dimer-inducing ligands, as well as those ligands that do not favor $Pd^{(I)}$ dimers on the Fey ligand map (*12*), it is apparent that the dimer-inducing ligands P(1-Ad)$_2$(*n*-Bu) and P(*t*-Bu)$_2$Ph are in fact closer to PCy_3, which does not favor $Pd^{(I)}$, than to the other dimer-inducing ligands P(*t*-Bu)$_3$ and P(*t*-Bu)$_2$(*i*-Pr) (Fig. 1C). Moreover, beyond these 2D representations, our analysis of the Euclidean distance of the ligands with the descriptors of this database (fig. S5) also categorized two of the dimer-inducing ligands [P(1-Ad)$_2$(*n*-Bu) and P(*t*-Bu)$_2$Ph] closer to PCy_3 than to the other dimer-inducing ligands.

Because existing qualitative guides and insight-driven strategies clearly fail for this (and other) speciation challenge(s), we set out to explore alternative means of chemical guidance and investigated the feasibility of a data-driven approach in this context.

Among machine learning methods, so called "supervised" and "unsupervised" algorithms represent the most common forms of learning (*22*). In supervised learning, models are trained with data consisting of input-output pairs. Supervised learning is especially useful for regression and classification tasks and has been successfully applied to predict selectivities (*23*–*26*), as well as reaction conditions (*27*, *28*) and yields (*29*, *30*) in the context of catalysis (*31*–*34*). However, large training datasets are needed as a prerequisite for this approach,

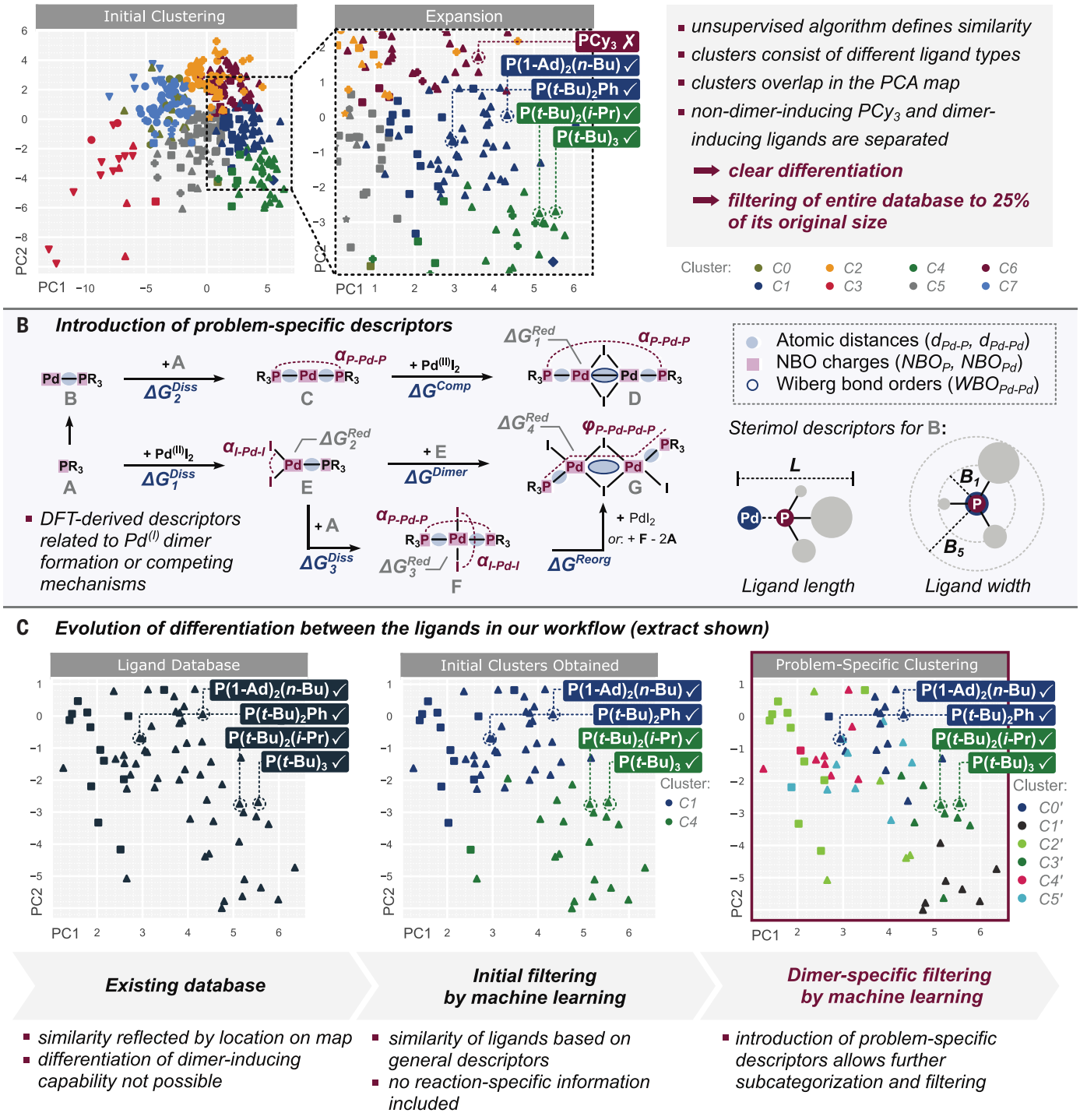

Fig. 2. Data generation and clustering. (**A**) Initial clustering of the LKB-P using the unsupervised *k*-means algorithm (*k* = 8; see figs. S8 and S9 for further visualization). (**B**) Newly introduced descriptors relating to $Pd^{(I)}$-dimer formation (see table S2 for details). (**C**) Illustration of the differentiation of ligands after initial clustering (middle) and second, problem-specific refinement (right; see fig. S16 for detailed plot) versus original database (left). The same subset of ligands is illustrated.

which are not available in our (and many other) speciation challenge(s).

By contrast, unsupervised machine learning techniques can be applied to recognize patterns in datasets without requiring a training of the algorithm with labeled data (and therefore without the known outputs, such as experiments). The learning process provides insights that are fundamentally different from traditional analyses, as they are derived purely by the "machine" without "human" guidance. Clustering is one of the main fields of unsupervised learning, whereby data are divided into several groups (clusters) with respect to the underlying similarity of the data points.

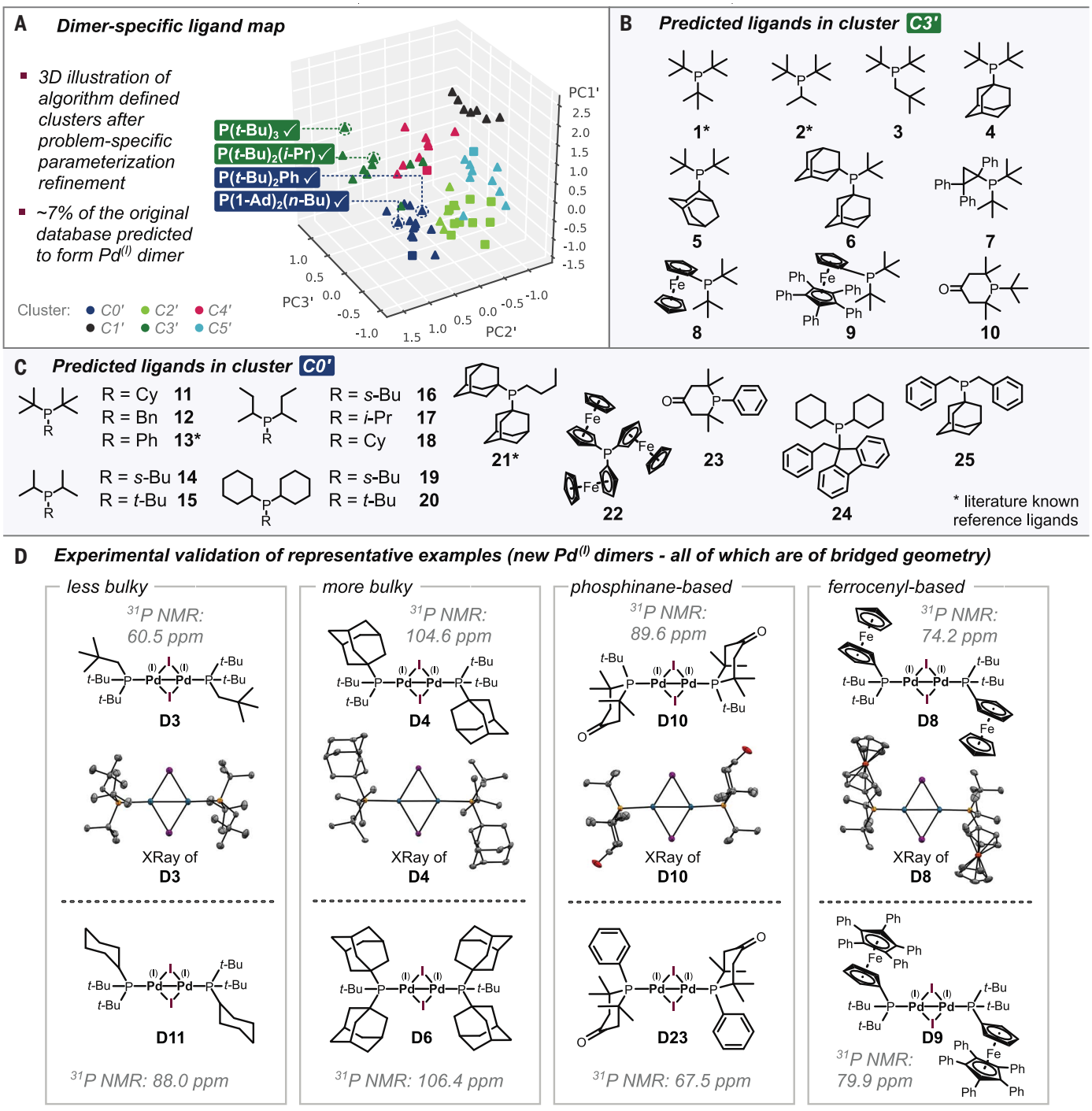

Fig. 3. Predictions and experimental tests. (**A**) Results of the problem-specific clustering—3D illustration (see figs. S17 to S19 for 2D plots and movie S1 for a 3D animation). (**B** and **C**) Ligands that the algorithm grouped into clusters C3′ and C0′ (see fig. S20 for cluster association of all ligands). (**D**) Experimental testing: newly synthesized $Pd^{(I)}$ dimers and their x-ray crystallographic structures and representative ^{31}P NMR shifts (H atoms removed for clarity in the x-ray images).

Initial clustering of ligand knowledge base

The starting point for our data analysis was the ligand knowledge base for monodentate P-donor ligands (LKB-P) published in 2010, which covers a total of 348 ligands featuring eight different ligand types, such as various phosphines, phosphites, and other hetero-atom-containing variants (*11*, *12*). Roughly 30% of the ligands in this database are not commercially available and 17% do not have published experimental syntheses (according to our Scifinder search in May 2021).

We set out to add another layer of knowledge to the original LKB-P database by using unsupervised machine learning to examine whether the algorithm would detect distinct relations between the ligands (Fig. 2A). We chose the *k*-means algorithm (see section S1 in the supplementary materials) to partition the ligands into different subgroups (clusters). This algorithm requires only one parameter

to be defined in advance—i.e., the so called "number of clusters *k*," for which we chose a value of $k = 8$ (based on the analysis of the elbow method and silhouette scores) (*13*).

The algorithm clustered the ligands independent of their actual ligand type, often mixing different types within the same cluster in a nonintuitive way (data S1). Moreover, although ligands were in the same area of the Fey ligand map, they were assigned to different clusters, making the clusters overlap, especially in the top part of the map (Fig. 2A). This overlap suggests that the algorithm detected certain differences between the ligands, which the purely visual analysis of the original ligand map (Figs. 1C and 2C) would not have revealed. In particular, the algorithm differentiated the known four $[Pd^{(I)}(\mu\text{-}X)L]_2$-dimer–inducing ligands—i.e., P(*t*-Bu)$_3$, P(*t*-Bu)$_2$(*i*-Pr), P(*t*-Bu)$_2$Ph, and P(1-Ad)$_2$(*n*-Bu)—from the non–dimer-inducing PCy_3. Whereas the dimer-inducing ligands were grouped within the clusters C1 and C4 (blue and green; Fig. 2A), the non–dimer-inducing PCy_3 was part of a separate cluster (C6, dark red); this outcome was verified by reperforming the clustering 1000 times with different (random) initialization seeds (*13*). These findings motivated us to further investigate the 89 other ligands that were also assigned to the clusters C1 and C4.

Subjection to another *k*-means–based clustering for further subcategorization of the predicted ligands using the same descriptors (which was performed 1000 times with different initialization seeds for a statistically relevant picture) resulted in insufficient reduction of ligand space however, with 60% of these ligands being grouped in the same clusters as those of our four known dimer-inducing reference ligands.

Because the descriptors of the original LKB-P database were selected to describe the ligands as generally as possible, the 89 ligands can be regarded as similar in terms of their general properties, which is likely why the attempt to further subcategorize was unproductive. However, to differentiate these ligands for our specific chemical question at hand—i.e., whether a ligand will favor the dihalide-bridged $Pd^{(I)}$ dimer or not—we envisioned that a further refinement of the clusters C1 and C4 toward that criterion would be optimal.

Owing to the first clustering, where all ligands were grouped on the basis of their general properties, the ligand space had been substantially reduced from 348 to 89 ligands (i.e., to ~25% of the original database). The problem-specific data therefore only need to be generated for this subset. With this approach, we essentially guide the algorithm to examine a certain subset of chemical ligand space more closely for our problem at hand, while minimizing the need for additional data generation.

Introducing problem-specific descriptors

We set out to introduce a set of new problem-related descriptors, which we obtained entirely in silico through quantum-mechanical [density functional theory (DFT)] calculations. We decided to focus exclusively on phosphine ligands that contained P-C bonds, comprising a total of 66 ligands (see fig. S20 for an illustration). Following the workflow depicted in Fig. 1D, we introduced a set of new descriptors derived from DFT calculations on specific aspects relating to $Pd^{(I)}$ dimer formation (Fig. 2B). We included mono- and bisphosphine $Pd^{(0)}$ and $Pd^{(II)}$ complexes (species B, C and E, F in Fig. 2B), as well as the dinuclear $Pd^{(I)}$ and potentially competing dinuclear $Pd^{(II)}$ complexes (species D and G).

In this context, we focused solely on iodide-bridged $Pd^{(I)}$ entities as those complexes tend to be air stable (*15*). We preserved the information from the initial clustering in our refined model by keeping some of the general descriptors from the initial dataset, such as the HOMO/LUMO energies, the proton affinity, and the measure of steric bulk of the phosphines, as well as the Pd-related descriptors (*10*). Subsequently, we derived a total of 42 new descriptors to represent the effects of the phosphines in the dimer-related context. Aiming to capture general Pd-ligand bond properties, we introduced various geometrical descriptors for all calculated complexes (Fig. 2B). In addition, Sterimol descriptors (*35*) were introduced to capture the impact of conformational effects. Electronic descriptors included natural bond orbital (NBO) charges of palladium (NBO_{Pd}) and phosphorus centers (NBO_P) (Fig. 2B). Free energies of reduction ($\Delta G^{Red}_1 - \Delta G^{Red}_4$) were calculated for the species D to G (via the hypothetical reduction that yields the corresponding neutral $Pd^{(0)}$ species and iodide anions), as well as ligand bond-dissociation energies ($\Delta G^{Diss}_1 - \Delta G^{Diss}_4$) for C, E, and F. Moreover, the comproportionation energies (ΔG^{Comp}) for the formation of $Pd^{(I)}$ dimer (D) from bisphosphine $Pd^{(0)}$ (C) and PdI_2, as well as the dimerization energy (ΔG^{Dimer}) to form $Pd^{(II)}$ dimer (G) from either the corresponding monomers (E) or by reorganization (ΔG^{Reorg}) from the bisligated $Pd^{(II)}$ complex (F) and PdI_2, were included. Although the previous mechanistic interpretation of these reaction energies was not conclusive, the values were included in the clustering to monitor potential relative trends between the ligands. Pd-I-I-Pd torsion and Wiberg bond indices of the Pd-Pd bond ($WBO_{Pd\text{-}Pd}$) were used to describe the dinuclear complexes D and G and to capture stability trends. To assess the quality of these new descriptors, we analyzed their correlation using the absolute Pearson correlation coefficient (fig. S12) (*13*).

Principal component analysis was performed to obtain an improved visualization of the problem-specific ligand space and to investigate the contribution of the new descriptors to the variance of the data (see supplementary materials section S3.2.2 and fig. S13). Both approaches indicate the suitability of the newly introduced descriptors for further analysis and confirmed the substantial contribution of all descriptors to the new principal components.

Problem-specific clustering

Following the generation of new $Pd^{(I)}$ dimer-related data, we proceeded with the *k*-means clustering and chose $k = 6$ on the basis of the elbow method and silhouette score analyses (see section S3.3 in supplementary materials). We subsequently plotted the resulting six clusters in the space spanned by the principal components of the initial database (PC1 and PC2). Figure 2C (right) shows that the problem-specific refinement resulted in further subdifferentiation of the previously obtained two clusters (middle). Figure 3A gives an additional 3D illustration of the full dataset after problem-specific clustering. The three new principal components, PC1′, PC2′, and PC3′, capture 60.7% of the variation in the problem-specific data (*13*). Of these six visually separated clusters, two clusters included the known dimer-inducing ligands, with P(*t*-Bu)$_3$ and P(*t*-Bu)$_2$(*i*-Pr) in cluster C3′, whereas P(*t*-Bu)$_2$Ph and P(1-Ad)$_2$(*n*-Bu) were grouped in cluster C0′. The other members of these two clusters would therefore be expected to similarly favor dihalide-bridged $Pd^{(I)}$ dimers. Their similarity to the references was again verified by reperforming the clustering 1000 times with different initialization seeds.

Closer inspection of the other ligands that the algorithm grouped into C3′ and C0′ indicated that cluster C3′ contained several relatively bulky trialkyl phosphine ligands (**1** to **6**, Fig. 3B) closely resembling the already known dimer-inducing motif. The algorithm-identified similarity would hence largely also be in line with chemical intuition. Unexpectedly, however, and not necessarily in line with intuition, the remaining four ligands in the cluster C3′ differ structurally to an extent that a similarity is not clearly obvious. For example, the sterically constrained and rather bulky cyclopropyl derivative (cBRIDP **7**) and the two ferrocenyl-based phosphines (**8** and QPhos **9**) are included in the same group, although the electronic impacts of the latter differ considerably from those of the trialkyl phosphine series owing to the aromaticity of the ferrocenyl-substituent. Especially unexpected was inclusion of the phosphinane ligand (**10**), in which the ring constrains the methyl groups into fixed positions, resulting in decreased flexibility compared to the other alkyl-based phosphines in the cluster.

In cluster C0′ as well, the machine learning predicted similarity is neither intuitive nor obvious. The phosphine substituents in this cluster are more diverse compared to those in cluster C3′, showing tertiary, secondary, and

primary alkyl chains, as well as aryl and benzyl groups (Fig. 3C). Although PCy_3 does not favor $Pd^{(I)}$ dimers, the algorithm predicted five ligands with one or two Cy groups.

We next set out to experimentally test these predictions (see section S4 in the supplementary materials for details). Of the 25 ligands in clusters C3′ and C0′, the trialkylphosphine ligands **1**, **2**, and **21,** as well as the phenyl derivative **13**, were previously reported to form $Pd^{(I)}$ dimers (*15*) and so were used by us to evaluate the results obtained from clustering. Of the 21 remaining predicted ligands, the ferrocene-derivatives **8** and **9**, as well as the highly constrained phosphinanes **10** and **23**, deviated most from the other ligands. We therefore set out to initially test these ligands for their dimer-inducing capability. Although the ferrocenyl ligands have found numerous applications in typical (presumed) $Pd^{(0)}/Pd^{(II)}$-catalyzed cross-coupling applications (*36*), the phosphinane **10** has no reported synthesis, and there is hence also no known application of it. We attempted the synthesis of the corresponding iodide-bridged $Pd^{(I)}$ dimers by comproportionation of $Pd^{(0)}L_n$ and PdI_2, following the syntheses of the corresponding $Pd^{(0)}L_n$ complexes (and the ligands, respectively).

Finally, we observed clean formation of the $Pd^{(I)}$ dimers **D8** and **D10** as supported by ^{31}P nuclear magnetic resonance (NMR) and x-ray crystallographic analysis (Fig. 3D). Also, our tests of the adamantyl-rich ligands **4** and **6**, as well as the cyclohexyl-containing **11** and the neopentyl analog **3**, all successfully gave rise to the corresponding iodide-bridged $Pd^{(I)}$ dimers, as predicted by the algorithm. By contrast, for the very bulky cyclopropyl derived ligand **7**, we did not succeed in making a $Pd^{(I)}$ species. All our attempts to synthesize a $Pd^{(0)}L_2$ complex for subsequent comproportionation failed for this ligand; an alternative route via a $Pd^{(I)}$-$Pd^{(I)}$ template (*37*) was also not efficient, which is likely due to the ligand's size and diminished binding capacity to displace the precursor ligands.

With regard to the 41 ligands that were grouped in the four other clusters after the dimer-specific clustering (see fig. S20 for the structures), we also tested representative ligands from each group by mixing the ligand with $Pd_2(dba)_3$ (for in situ generation of $Pd^{(0)}L_2$) and PdI_2 in tetrahydrofuran and examined the mixtures by ^{31}P NMR. The ligands in C2′ (15 in total) are relatively aryl-rich, and our examination indicated that $Pd^{(II)}$ species likely resulted in these cases (see fig. S17). Clusters C1′ and C5′ contain predominantly biaryl ligands (17 overall), which can form $Pd^{(I)}$ dimers (*37*), albeit not in our targeted dihalide-bridged geometry but instead in an alternative cationic geometry, where the π-system acts as a bridge. The algorithm therefore appears to correctly discriminate the ligands not only in their ability to favor oxidation state (I) but also in favoring the desired dihalide-bridged dimer geometry. Lastly, C4′ (9 ligands) consists of phosphaadamantane ligands and also trialkyl phosphines, for which we observed a mixture of species in our test of a representative.

In the initial clustering of the 348 ligands of the entire database, we had eliminated six of the eight clusters, as only two contained dimer-inducing reference ligands. In this context, if we had only had two experimental data points as prior knowledge, e.g. PCy_3 (as non–dimer-inducing) and $P(t\text{-}Bu)_3$ (as dimer-inducing), we would have only selected a single cluster for further dimer-specific filtering (i.e., the one that contained $P(t\text{-}Bu)_3$) and eventually obtained fewer suggestions after dimer-specific clustering. As such, the extent of identification of dimer-inducing ligands naturally depends on the size of the initial database and the number of reference ligands to guide cluster selection. However, the scientist has the opportunity to experimentally examine representatives from each of the originally obtained eight clusters to identify another dimer-inducing cluster. This approach therefore confronts the vast potential chemical space in a highly efficient manner with minimal experimentation.

By reperforming the clustering 1000 times with different initializations, we eliminated the possibility that our predictions result from random fluctuations. Our 21 final ligand candidates were grouped with the four dimer-inducing reference ligands in ≥80% of all clusterings (four ligands in ~80%, 17 ligands in >90% of all clusterings). There were eight additional ligands encountered in the 1000 clusterings in fewer cases—i.e., one ligand in 50%, one in 40%, and all others in < 30% of all clusterings. In accord with their low score, our experimental tests of three of these [that were readily synthetically accessible, i.e., $PMe(t\text{-}Bu)_2$, $P(t\text{-}Bu)Np_2$, and $PPh(s\text{-}Bu)_2$], confirmed that these ligands do not give iodide-bridged $Pd^{(I)}$ dimers. These results suggest that such an abundance value averaged over 1000 clusterings is a useful means for selection. As a further validation of its usefulness, we also examined the above-mentioned "failed" clustering without problem-specific data once again. In this case, we had subjected the results from the initial clustering of the LKB-P to another clustering with the same generalized descriptors (without introducing problem-specific data). This had led to insufficient filtering, as discussed above. In the 1000 clusterings we had performed in this context, there were six ligands predicted that were grouped with the four dimer-inducing reference ligands in >95% of all clusterings. However, our experimental tests of four of these six ligands indicated that these lead to $Pd^{(II)}$ species instead (figs. S20 and S23). These results reinforce the importance and success of introducing problem-specific descriptors for the second clustering and underline the success of our developed workflow.

Overall, we were able to experimentally verify numerous representative examples of the 21 predicted ligands and synthesized in total eight previously unreported air-stable $Pd^{(I)}$ dimers. As such, the algorithm was highly successful in recognizing similarities between the ligands that are not obvious to the human specialist's eye. Especially ligand **10** (which had never been made) would not likely have been investigated on a trial-and-error, screening or intuition-guided study. This is a clear demonstration of the power of machine learning techniques to accelerate catalyst development with suggestions that are beyond a scientist's intuition. Our future efforts are directed at exploring the potential of the new dimers in catalysis (*38*, *39*).

REFERENCES AND NOTES

1. J. F. Hartwig, *Organotransition Metal Chemistry: From Bonding to Catalysis* (University Science Books, New York, 2010).
2. F. Barrios-Landeros, B. P. Carrow, J. F. Hartwig, *J. Am. Chem. Soc.* **130**, 5842–5843 (2008).
3. C. J. Diehl, T. Scattolin, U. Englert, F. Schoenebeck, *Angew. Chem. Int. Ed.* **58**, 211–215 (2019).
4. M.-E. Moret, R. J. M. Gebbink, Non-Noble Metal Catalysis: Molecular Approaches and Reactions (Wiley-VCH, Weinheim, Germany, 2019).
5. C. A. Tolman, *Chem. Rev.* **77**, 313–348 (1977).
6. W. Strohmeier, F.-J. Müller, *Chem. Ber.* **100**, 2812–2821 (1967).
7. C. A. Tolman, *J. Am. Chem. Soc.* **92**, 2953–2956 (1970).
8. K. A. Bunten, L. Chen, A. L. Fernandez, A. J. Poë, *Coord. Chem. Rev.* **233-234**, 41–51 (2002).
9. Z. L. Niemeyer, A. Milo, D. P. Hickey, M. S. Sigman, *Nat. Chem.* **8**, 610–617 (2016).
10. D. J. Durand, N. Fey, *Chem. Rev.* **119**, 6561–6594 (2019).
11. N. Fey *et al*., *Chemistry* **12**, 291–302 (2005).
12. J. Jover *et al*., *Organometallics* **29**, 6245–6258 (2010).
13. See supplementary materials for additional details and definition of terminology.
14. C. C. C. Johansson Seechurn, M. O. Kitching, T. J. Colacot, V. Snieckus, *Angew. Chem. Int. Ed.* **51**, 5062–5085 (2012).
15. C. Fricke, T. Sperger, M. Mendel, F. Schoenebeck, *Angew. Chem. Int. Ed.* **60**, 3355–3366 (2021).
16. J. P. Stambuli, R. Kuwano, J. F. Hartwig, *Angew. Chem. Int. Ed.* **41**, 4746–4748 (2002).
17. D. M. Ohlmann *et al*., *J. Am. Chem. Soc.* **134**, 13716–13729 (2012).
18. X.-Y. Chen, M. Pu, H.-G. Cheng, T. Sperger, F. Schoenebeck, *Angew. Chem. Int. Ed.* **58**, 11395–11399 (2019).
19. S. T. Keaveney, G. Kundu, F. Schoenebeck, *Angew. Chem. Int. Ed.* **57**, 12573–12577 (2018).
20. Calculated at the CPCM (toluene) B3LYP-D3(BJ)/6-311++G(d,p)(SDD for Pd, Fe, I) //B3LYP-D3(BJ)/6-31G(d) (SDD for Pd, Fe, I) level of theory.
21. F. Proutiere, E. Lyngvi, M. Aufiero, I. A. Sanhueza, F. Schoenebeck, *Organometallics* **33**, 6879–6884 (2014).
22. J. B. O. Mitchell, *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **4**, 468–481 (2014).
23. J. P. Reid, M. S. Sigman, *Nature* **571**, 343–348 (2019).
24. A. R. Rosales *et al*., *Nat. Catal.* **2**, 41–45 (2019).
25. A. Milo, A. J. Neel, F. D. Toste, M. S. Sigman, *Science* **347**, 737–743 (2015).
26. A. F. Zahrt *et al*., *Science* **363**, eaau5631 (2019).
27. Y. Amar, A. M. Schweidtmann, P. Deutsch, L. Cao, A. Lapkin, *Chem. Sci.* **10**, 6697–6706 (2019).
28. H. Gao *et al*., *ACS Cent. Sci.* **4**, 1465–1476 (2018).
29. D. T. Ahneman, J. G. Estrada, S. Lin, S. D. Dreher, A. G. Doyle, *Science* **360**, 186–190 (2018).
30. C. W. Coley, R. Barzilay, T. S. Jaakkola, W. H. Green, K. F. Jensen, *ACS Cent. Sci.* **3**, 434–443 (2017).
31. M. Foscato, V. R. Jensen, *ACS Catal.* **10**, 2354–2377 (2020).
32. T. Toyao *et al*., *ACS Catal.* **10**, 2260–2297 (2020).

33. I. Funes-Ardoiz, F. Schoenebeck, *Chem* **6**, 1904–1913 (2020).
34. G. dos Passos Gomes, R. Pollice, A. Aspuru-Guzik, *Trends Chem.* **3**, 96–110 (2021).
35. A. V. Brethomé, S. P. Fletcher, R. S. Paton, *ACS Catal.* **9**, 2313–2323 (2019).
36. S. G. Newman, M. Lautens, *J. Am. Chem. Soc.* **133**, 1778–1780 (2011).
37. K. O. Kirlikovali *et al*., *Dalton Trans.* **47**, 3684–3688 (2018).
38. Ligand 20 that was identified in our study gave a successful dimer also which we meanwhile applied in olefin isomerization; see (*39*).
39. G. Kundu, T. Sperger, K. Rissanen, F. Schoenebeck, *Angew. Chem. Int. Ed.* **59**, 21930–21934 (2020).
40. J. A. Hueffel, J-Hueffel/PdDimer: PdDimer v1.0, *Zenodo* (2021); https://doi.org/10.5281/zenodo.5541842.

ACKNOWLEDGMENTS

Funding: This work was funded by the Volkswagen Foundation (Momentum Program), the Alexander von Humboldt Foundation (fellowship to I.F.-A.), and the DFG (German Research Foundation) Cluster of Excellence 2186 ("The Fuel Science Center" – ID: 390919832). Calculations were performed with computing resources granted by JARA-HPC from RWTH Aachen University under project "jara0091." **Author contributions:** J.A.H., T.S., and I.F.-A. performed the computational and data science research and T.S. the experiments. J.S.W., K.R., and T.S. solved the x-ray structures. J.A.H., T.S., I.F.-A., and F.S. analyzed the research data and contributed to the writing of the manuscript. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** ML algorithm and python code are available online at GitHub (*40*). Crystallographic data are available at the Cambridge Crystallographic Data Center (www.ccdc.cam.ac.uk) under reference numbers CCDC 2055171 (**D3**), 2064562 (**D4**), 2055172 (**D8**), 2055173 (**D10**), 2055174 (**D11**), and 2064863 (**D23**). All other data are available in the main text or the supplementary materials.

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.abj0999
Materials and Methods
Supplementary Text
Figs. S1 to S23
Tables S1 to S6
References (*41*–*80*)
Movie S1
Data S1 to S3

20 April 2021; accepted 30 September 2021
10.1126/science.abj0999

Science

Accelerated dinuclear palladium catalyst identification through unsupervised machine learning

Julian A. HueffelTheresa SpergerIgnacio Funes-ArdoizJas S. WardKari RissanenFranziska Schoenebeck

Science, 374 (6571), • DOI: 10.1126/science.abj0999

Learning to stabilize palladium dimers

Catalyst optimization is often difficult to do rationally. Once something works, it may be unclear which specific features underpin the performance. A case in point is the stabilization of palladium(I) dimers, which has relied on a very small class of phosphine ligands. Hueffel *et al*. used machine learning to search for patterns in this known class of ligands and thereby guide the discovery of variants that likewise stabilize the dimers. The authors were able to synthesize eight previously unreported dimers. —JSY

Science (ISSN) is published by the American Association for the Advancement of Science. 1200 New York Avenue NW, Washington, DC 20005. The title *Science* is a registered trademark of AAAS.