A comprehensive discovery platform for organophosphorus ligands for catalysis

Tobias Gensch*§[1,2], Gabriel dos Passos Gomes§[3,4,5], Pascal Friederich§[3,4,6], Ellyn Peters[1], Théophile Gaudin[4,7], Robert Pollice[3,4], Kjell Jorner[3,4,8], AkshatKumar Nigam[3,4], Michael Lindner-D'Addario[3,4], Matthew S. Sigman*[1], Alán Aspuru-Guzik*[3,4,5,9]

[1] Department of Chemistry, University of Utah, 315 South 1400 East, Salt Lake City, UT 84112, USA.
[2] Department of Chemistry, TU Berlin, Straße des 17. Juni 135, Sekr. C2, 10623 Berlin, Germany.
[3] Chemical Physics Theory Group, Department of Chemistry, University of Toronto, 80 St George St, Toronto, Ontario M5S 3H6, Canada.
[4] Department of Computer Science, University of Toronto, 214 College St., Toronto, Ontario M5T 3A1, Canada.
[5] Vector Institute for Artificial Intelligence, 661 University Ave Suite 710, Toronto, Ontario M5G 1M1, Canada.
[6] Institute of Nanotechnology, Karlsruhe Institute of Technology, Hermann-von-Helmholtz-Platz 1, 76344 Eggenstein-Leopoldshafen, Germany.
[7] IBM Research Zurich, Säumerstrasse 4, 8803 Rüschlikon, Switzerland.
[8] Early Chemical Development, Pharmaceutical Sciences, R&D, AstraZeneca, Macclesfield, United Kingdom.
[9] Lebovic Fellow, Canadian Institute for Advanced Research (CIFAR), 661 University Ave, Toronto, Ontario M5G
§ *These authors contributed equally to this work.*
* *Corresponding authors:* tobias.gensch@tu-berlin.de, matt.sigman@utah.edu, aspuru@utoronto.ca

Abstract

The design of molecular catalysts typically involves reconciling multiple conflicting property requirements, largely relying on human intuition and local structural searches. However, the vast number of potential catalysts requires pruning of the candidate space by efficient property prediction with quantitative structure-property relationships. Data-driven workflows embedded in a library of potential catalysts can be used to build predictive models for catalyst performance and serve as a blueprint for novel catalyst designs. Herein we introduce *kraken*, a discovery platform covering monodentate organophosphorus(III) ligands providing comprehensive physicochemical descriptors based on representative conformer ensembles. Using quantum-mechanical methods, we calculated descriptors for 1,558 ligands, including commercially available examples, and trained machine learning models to predict properties of over 300,000 new ligands. We demonstrate the application of *kraken* to systematically explore the property space of organophosphorus ligands and how existing datasets in catalysis can be used to accelerate ligand selection during reaction optimization.

Introduction

Ligand engineering on the basis of mechanistic hypotheses has been a primary driver of reaction discovery and optimization in catalysis. An emerging and complementary approach applies data-driven methods to molecular design by capturing multidimensional property relationships that directly influence performance.[1] The success of such data-driven approaches relies on the availability of powerful molecular representations[2–4] that can be used in a wide range of machine learning (ML) methods.[5–10] Organophosphorous (III) ligands are amongst the most widely-used ligands in homogeneous catalysis. In this study, we establish a comprehensive workflow to study these ubiquitous compounds that can be further extended to other ligand classes. The platform that we developed can be employed for inverse design of novel homogeneous catalysts inspired by past work in the context of molecular and materials discovery. For example, the Materials Project,[11] OQMD[12] and AFLOW[13] are tools for exploring the inorganic compound space that include databases, computer scripts for feature extraction, and ML toolkits. Additionally, the Harvard Clean Energy Project[14] has similar goals in the space of organic photovoltaics. Moreover, in the case of heterogeneous catalysis, the Catalysis-Hub[15] contains computed heterogeneous

reaction energies and the associated barriers, and the Open Catalyst Project[16] provides density functional theory (DFT) geometry relaxations for material surfaces with adsorbates. These illustrative examples provide the foundation for how our teams approached the development of a workflow for the chemical space of organophosphorus ligands.

In this context, Tolman introduced what experimentally measured descriptors now referred to as the Tolman electronic parameter (TEP)[17] and the Tolman cone angle[18] in order to quantify and rationalize phosphorus ligand properties over fifty years ago. These molecular descriptors allowed mapping of the phosphine property space and provided a tool to understand systematic trends in reactivity and stability using linear free energy relationships and substituent additivity approaches.[19,20] Building on this, the ligand knowledge bases (LKB) developed by Fey and coworkers[21–23] marked an impressive milestone in the mapping of ligand space. The LKB-P consists of computed properties of 366 monodentate organophosphorus ligands in typical coordination environments.[23] Recently, the profound impact of ligand dynamics on catalytic reactions has been recognized more frequently,[24,25] however, the systematic quantification of ligand flexibility is still underdeveloped. Thus, inspired by the previous approaches to the mapping of ligand space, we aimed at developing a workflow that encompasses a wide range of steric and electronic properties of catalytically relevant ligands including descriptors for their flexibility to enhance the capabilities of data-driven catalyst design.[26]

Herein we present *kraken,* an extensive virtual open-access library covering monodentate organophosphorus(III) ligands targeted at facilitating the design and optimization of catalytic processes (**Figure 1**). To account for conformational flexibility, a general-purpose physicochemical descriptor set is derived from representative conformer ensembles of both the uncoordinated ligands and a model complex, which we hypothesized provides access to the essential features describing the multitude of intermolecular interactions involved in catalytically relevant steps. Additionally, we demonstrate the application of *kraken* to explore the property space of organophosphorus(III) ligands at a massive scale using increasingly sophisticated models for property estimation of arbitrary organophosphorus(III) compounds. Finally, we showcase the use of *kraken* for inverse catalyst design by constructing multiple linear free energy relationships based on experimental data and using it to predict the performance of the entire ligand database, providing the best candidates to be tested in subsequent experiments.

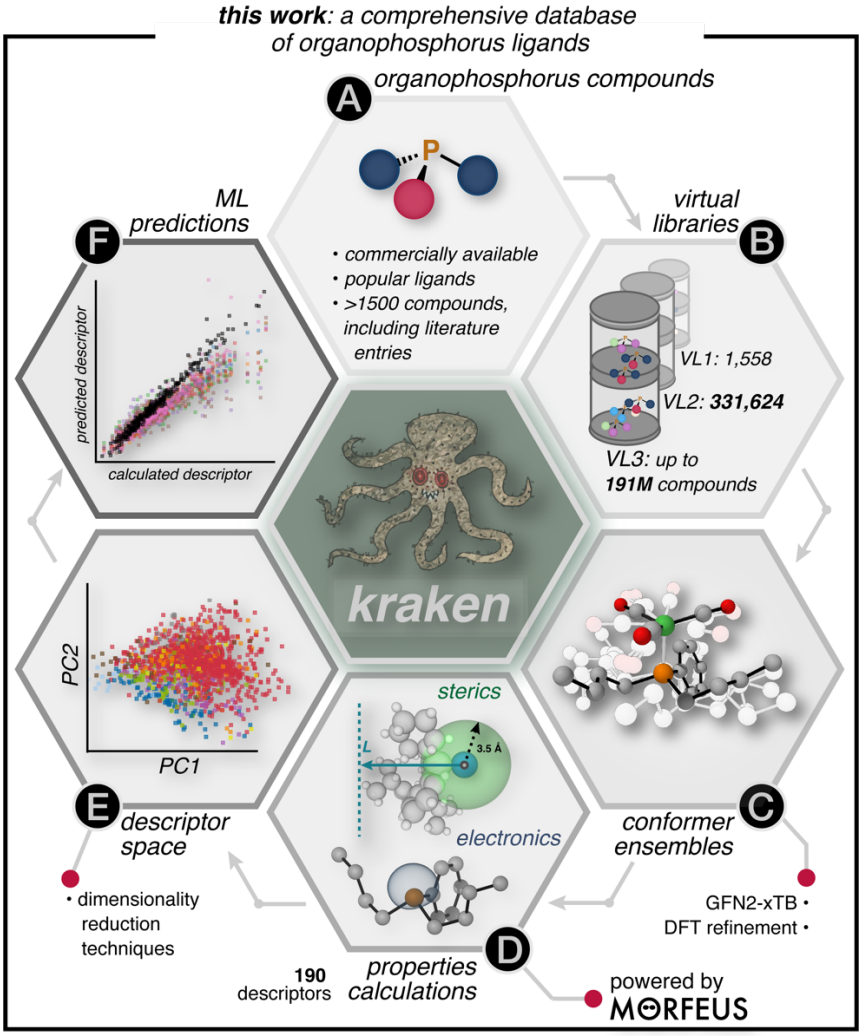

Figure 1: *kraken*: a comprehensive database of organophosphorus ligands. A. A set of 1558 organophosphorus compounds is gathered, including literature and commercial sources. **B.** Virtual libraries (VL) are built from the substituents of the initial P(III) set. The first level (Virtual Library 1, VL1) contains the initial set; VL2 results from a combinatorial approach with two different substituents per ligand (576 total unique fragments), yielding 331,776 compounds; VL3 is a virtual library where all combinations are possible, i.e., all 3 substituents can be different, with over 191 million entries. **C.** Conformer ensembles are generated for each of the P(III) molecules in VL1, at the GFN2-xTB level of theory. Each conformer is reoptimized using DFT, with a total of 21,437 conformers evaluated (average of 13.8 conformers per ligand). **D.** 78 physical organic properties are captured for every calculated conformer; Boltzmann averages, min-max steric extrema, and other representative conformers are curated for a total of 190 descriptors per ligand. **E.** Chemical property spaces are defined and visualized using dimensionality reduction techniques. **F.** ML models are built to simulate a virtual property library for approximately 330,000 compounds in VL2. VL3 is deployed by querying the ML models on demand.

Results and Discussion

Library Scope

A central goal was to comprehensively map the chemical space of monodentate organophosphorus(III) ligands, focusing in particular on structures relevant to applications in catalysis and its use for data-driven ligand optimization campaigns. We initially selected phosphines that were

commercially available and prevalent in the organo(transition)metal chemistry literature. In anticipation of the ML property prediction goals, we surveyed the scientific literature and systematically added ligands with less prevalent substituents based on the core structures found. This was followed by a curation step to avoid structures with additional N, P, or S-containing donor sites, and any acidic moieties as these structures may bind to the metal through other modes (e.g., bidentate ligands). Overall, the library contains ligands with various phosphorus-element bonds encompassing H, B, C, N, O, F, Si, and S next to phosphorus in arbitrary combinations. Thus, besides phosphines, other important ligand classes such as phosphoramidites, phosphites, and phosphinamines are also included. In its current state, library VL1 (cf. **Figure 1**) is constructed on full DFT calculations for 1,558 ligands and their conformers, at least 400 of which are commercially available and including the 200 most-cited phosphorus ligands in the literature.

Conformer Ensembles

One key challenge when defining the *kraken* computational workflow is the representation of the conformational space of each ligand, the conformer energies, and the corresponding contribution to the ligand properties. This is particularly relevant for steric properties that vary significantly with conformation, whereas electronic properties are generally less sensitive.[27] While no individual model system (i.e., free ligand or specific reference complexes) can fully reflect the conformational space accessible to a ligand in any given complex, there are certain limits for attainable geometries and properties. Importantly, investigating these ranges and limits was used to probe the behavior of ligands in catalytic systems and predict their catalytic performance. For example, the buried volume, i.e., the fraction of the volume of a sphere, which is placed at the metal center, occupied by ligand atoms,[28] of a trialkyl phosphine could be very large if all chains are folded towards the phosphorus lone pair, but it could never be smaller than when all chains are folded away (cf. **Figure 2C**). Thus, the smallest and largest attainable property values within the thermally accessible conformers of each ligand is defined as the representative range, irrespective of the exact complex environment. Notably, the correct range can only be derived from a (sufficiently) complete conformer ensemble. To allow the workflow to operate at large scale and reasonable cost, we applied GFN2-xTB,[29,30] a semiempirical tight-binding method developed to deliver excellent molecular geometries at the fraction of the cost of DFT, together with the workflows implemented in CREST to generate conformer ensembles.[31,32] Because of the sensitivity of steric properties to structural changes, we used these ensembles to select the structures with extreme values for at least one steric descriptor to be evaluated using DFT. Importantly, the conformational space of each ligand was assessed in two reference states, free ligand and coordinated to $Ni(CO)_3$. Generally, conformations in the free ligand tend to occupy more space around the phosphorus lone pair and, hence, free ligands appear more sterically demanding than complexed ones. Both situations are important to describing catalytic processes including potential unwanted side reactions like ligand dissociation. For consistent results, the ligand conformations from both reference states are then optimized as free ligands using DFT. To distinguish between a ligand and its individual conformers, we ascribe *properties* to the individual conformers of a ligand and *descriptors* to a ligand.

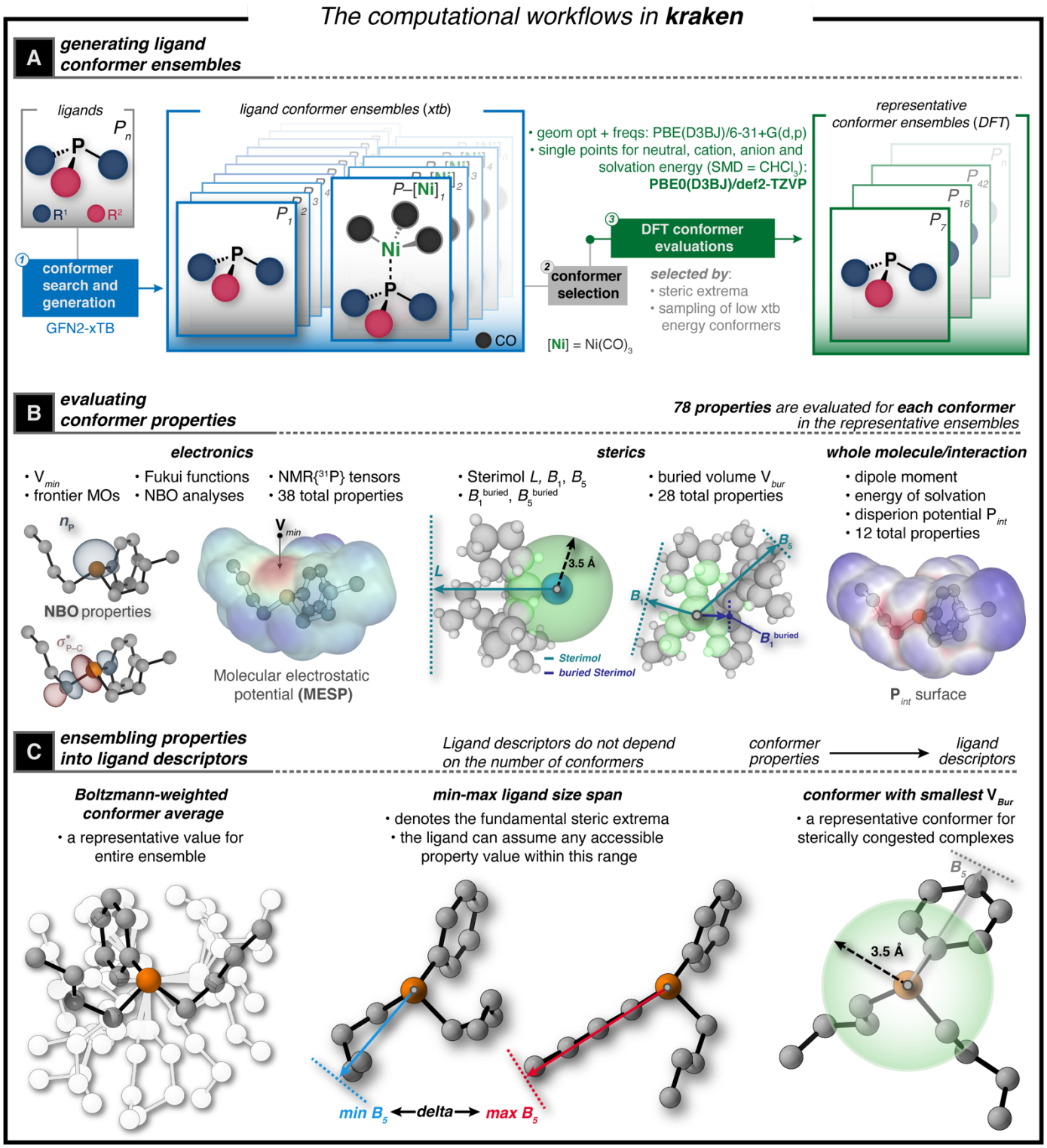

Figure 2: The computational workflows used to build *kraken*. **A**. Free and $Ni(CO)_3$-complexed ligands from VL1 are subjected to a conformer search with CREST at the GFN2-xTB level. Ligand conformer ensembles are subjected to a conformer selection. DFT is used for geometry optimization and single points of the selected conformers as free ligands. **B**. Illustrations of some properties computed for each conformer. **C**. Ensembling conformer properties to generate ligand descriptors.

Chemical Space Analysis

With this dataset, we set out to map the associated property space, understand the corresponding property limits and unveil uncharted regions potentially inspiring forays towards new unique ligand classes. The traditional analysis of phosphine properties uses Tolman's steric and electronic map, with the TEP on the abscissa and the Tolman cone angle on the ordinate.[33] This simple yet powerful visualization technique has helped chemists to survey available ligands rapidly and select structures with appropriate steric and electronic properties for specific applications. A more sophisticated version of Tolman's map has been introduced by Fey and co-workers using LKB[21,23] (see above) by reducing multiple descriptors to two dimensions via principal component analysis (PCA). Inspired by this work, we applied the recently developed Uniform Manifold Approximation and Projection[34] (UMAP, **Figure 3A**) as well as PCA (**Figure 3B**) to our entire database of DFT-computed ligands with all computed descriptors. These dimensionality reduction representations are available to interrogate on the interactive web application (https://kraken.cs.toronto.edu).

Nonlinear dimensionality reduction techniques can be employed to cluster compounds with a similar distribution of properties and for segregating distinct ligand classes from each other. For this purpose, we applied the UMAP technique as it preserves both local and global structure in the data and is computationally efficient.[34] The corresponding result is shown in **Figure 3A** using the elements bonded directly to phosphorus as a color code to illustrate the major phosphorus ligand classes. It is immediately obvious that the various ligand classes are well separated demonstrating the superior ability for data classification of UMAP. This suggested that our descriptor set contains relevant information to differentiate chemically distinct ligand types. Notably, UMAP essentially segregates the database into two important ligand superclasses, i.e., phosphorus bound to relatively electropositive elements like carbon and silicon and phosphorus bound to at least one relatively electronegative element like oxygen or nitrogen, with some overlaps between these two. Importantly, this aligns well with the binding affinities of these ligands as the atom type bound to phosphorus affects this property most severely.

The principal components obtained from PCA define a linearly uncorrelated descriptor set condensing the information contained in the database to as few dimensions as possible, while approximately preserving distance information in the descriptor space. This preservation of distances allows us to interpolate linearly between points in the descriptor space and, hence, understand the properties of unexplored regions as well. Accordingly, the resulting first two principal components were used to visualize the property space as depicted in **Figure 3B**. Again, by coloring the data points by the corresponding elements attached to the phosphorus atom, we can explore the relationships between common ligand classes, such as a smooth transition from phosphines (red) to phosphites (blue) via the intermediate phosphinites and phosphonites (purple) in the lower left of the chemical space.

Furthermore, not only can various ligand classes be distinguished, but the resulting principal components can be analyzed with respect to the properties they are encoding. PC1 generally represents total volume and PC2 pyramidalization. A more detailed analysis can be found in the ESI. Nevertheless, since the PCs combine various descriptors simultaneously, they represent a more integrated representation of the ligand space. Evaluating the next most heavily weighted principal components, PC3 is mainly determined by flexibility descriptors related to the inclusion of conformer ensemble property information and PC4 contains general orbital descriptors. Importantly, the added information from the computationally derived properties incorporates both depth and precision to compound representation as compared to Tolman's mapping. To provide a more intuitive illustration of the PCA property mapping, the

individual data points on the PCA plots were colored with respect to the buried volume[28] (V_{bur}, **Figure 3C**) and the minimum molecular electrostatic potential (MESP) in the phosphorus lone pair region,[35] which is correlated with the experimentally determined TEP (V_{min}, **Figure 3D**). Notably, these plots demonstrate that PC1 generally trends with V_{bur} and PC2 trends with V_{min}, even though it is not strongly collinear.

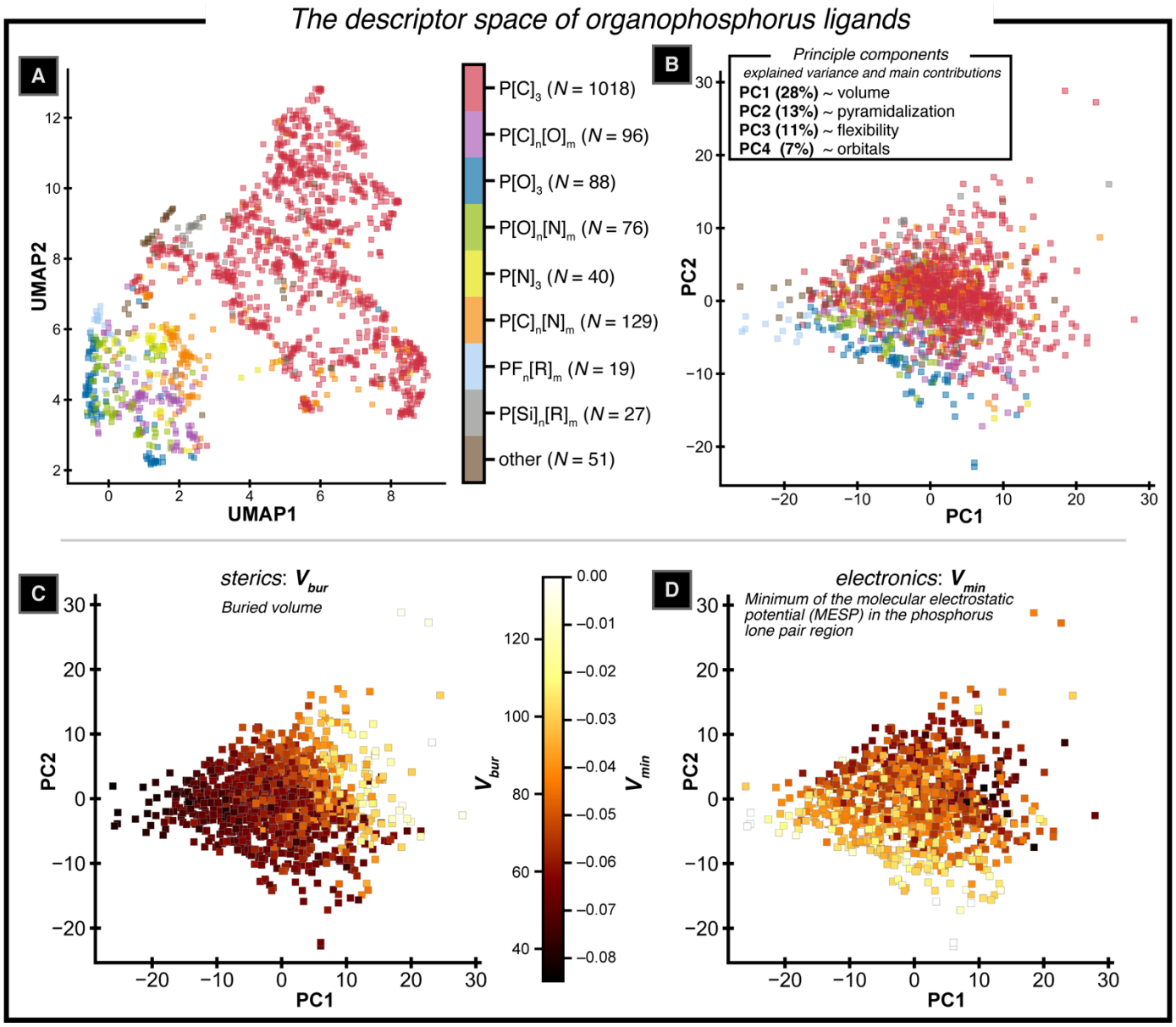

Figure 3: **Properties space visualization of monodentate organophosphorus(III) ligands using UMAP and PCA**. **A**. Dimensionality reduction of the descriptor space with UMAP to two dimensions. **B**. Dimensionality reduction of the descriptor space with PCA and projecting the corresponding results onto the two largest principal components, PC1 and PC2. **C**. Data points projected onto PC1 and PC2 subsequently colored by Boltzmann-averaged V_{bur}. **D**. Data points projected onto PC1 and PC2 later by Boltzmann-averaged V_{min}.

It is envisioned that these property maps can be used intuitively by process chemists that may not be experts in data science. Specifically, when the basic requirements in terms of sterics and electronics are known from previous experiments, the rational selection of the best ligand types that meet various process needs such as cost, environmental, and/or performance goals should be straightforward, similar to how the Solvent Selection Tool is applied by process chemists to locate the best solvent for a given reaction.[36]

Expanding the Space with Machine Learning

While we achieved a substantial coverage of the organophosphorus ligand space using quantum chemical simulations, 1558 compounds merely constitute a fraction of the conceivable space of this ligand class. Our computational workflow is too resource-intensive to probe all possible compounds of interest and explore the sparsely covered territory more comprehensively (see **Figure 3** and **Figure 4A**). Hence, to complement the simulations described above, we investigated several complementary ML methods to expand the compound space in our library significantly, and provide descriptor estimates for >300,000 molecules.

Inspired by the Benson group-increment theory[20,37] in thermochemistry and the demonstration of substituent additivity for the TEP by Tolman,[17] we tested if descriptors can be expressed as the sum of constant contributions from each substituent at phosphorus. To accomplish this, we represented each ligand as a matrix of all unique substituents bound to the central phosphorus atom containing the number of each substituent present in a particular compound (**Figure 4B**), which we term "Bag of Substituents" (BoS). For instance, PMe_2tBu would be encoded by the features "Me" and "*t*Bu", with a value of 2 in the former column, a value of 1 in the latter, and zeros in all other feature columns (576 in total). Linear regression of each descriptor individually was used with the BoS encoding to assess the additivity hypothesis. The coefficients of determination are a measure of how well the additivity assumptions hold for a descriptor and the trained weights correspond to the group increments. It should be noted that this model is inherently incapable of extrapolating to unseen substituents. As a consequence, all substituents needed to be included at least once in the training data, and, when possible, we enforced it to contain at least two occurrences. Apart from this constraint, we used a random 60:20:20 train-validation-test split. Good prediction quality was observed for a number of descriptors (58 properties with $R^2_{test} \geq 0.80$). As expected, $V_{min}^{(Boltz)}$ ($R^2 = 0.97$, cf. **Figure 5**) and $V_{bur}^{(Boltz)}$ ($R^2 = 0.95$) are well predicted. Interestingly, several descriptors that may not be expected *a priori* to be "additive" are also predicted with good accuracy, such as the Boltzmann-averaged NBO partial charge at the phosphorus atom ($R^2_{test} > 0.99$).

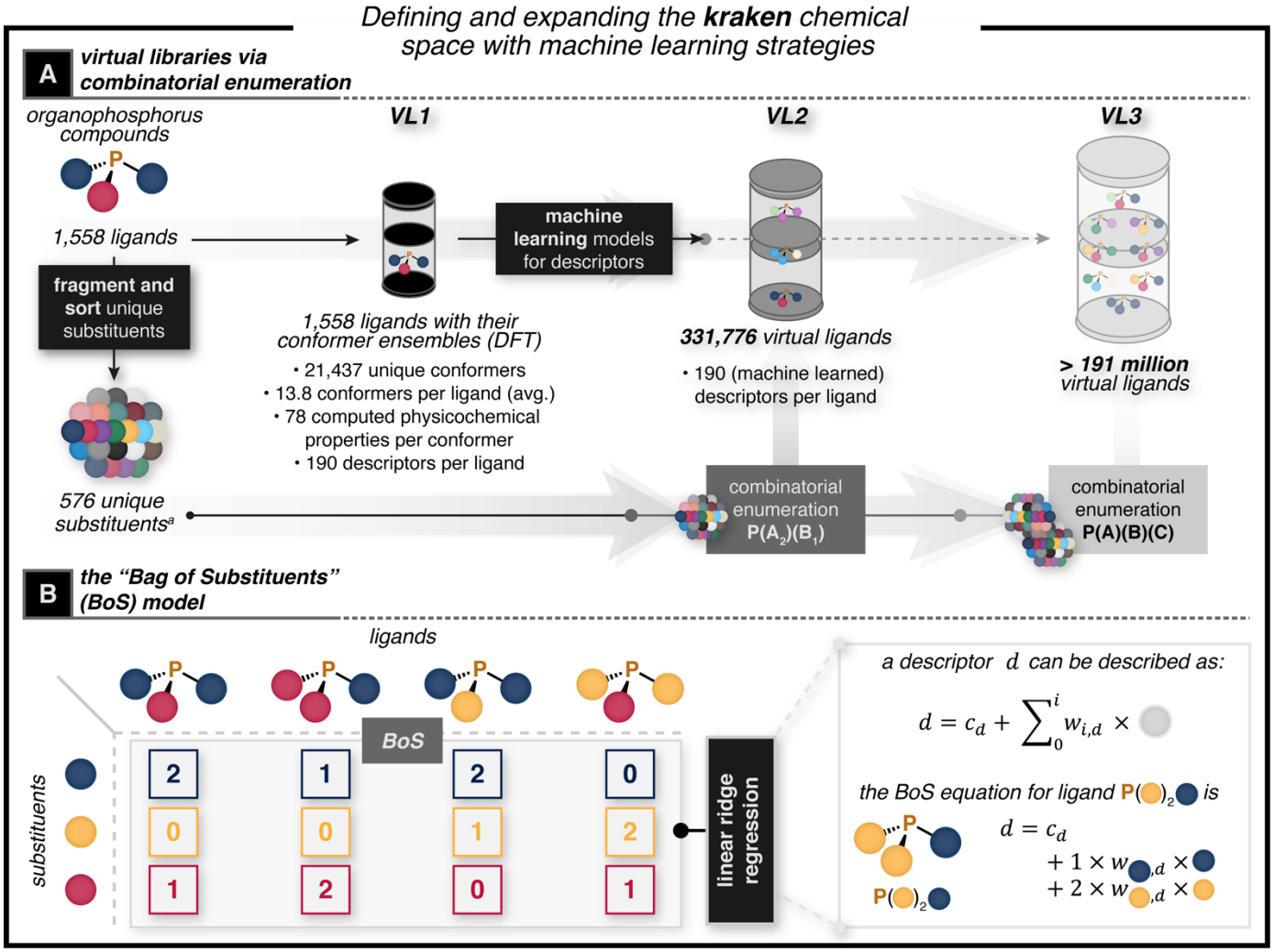

Figure 4: **Defining and expanding the *kraken* chemical space with machine learning strategies. A**. Construction of the virtual libraries VL2 and VL3, respectively, from virtual library VL1, which comprises 1558 unique ligands. [a]: the number of unique ligands excludes ligands in which the phosphorus atom is within a ring. **B**. Illustration of the "Bag of Substituents" model to predict ligand descriptors based on substituent increments. *d*: descriptor; c_d: constant; $w_{i,d}$: substituent weight per descriptor; *i* in the sum: total number of occurrences for a given substituent in a ligand.

While the BoS encoding strategy is relatively effective, some descriptors are not well predicted (45 properties with $R^2 < 0.50$) as is expected when substituent interactions or conformational effects are present that this simple model cannot incorporate. Thus, we used molecular fingerprints and graphs as more generalizable features to expand our predictive capacities. With those representations, we also applied other model types such as random forest (RF),[5] gradient boosting regressions (GBR),[38,39] Gaussian processes (GP),[39,40] and graph convolutional neural networks[41] (GCN, see **Figure 5** for the performance on one representative descriptor, more details on the ML models are in the ESI). Each of the models was found to be accurately predictive for a subset of descriptors. However, as none of the approaches were consistently the best for all the descriptors considered, we generated one metamodel for each descriptor. This was accomplished by ensembling all the models linearly to maximize the overall prediction quality.

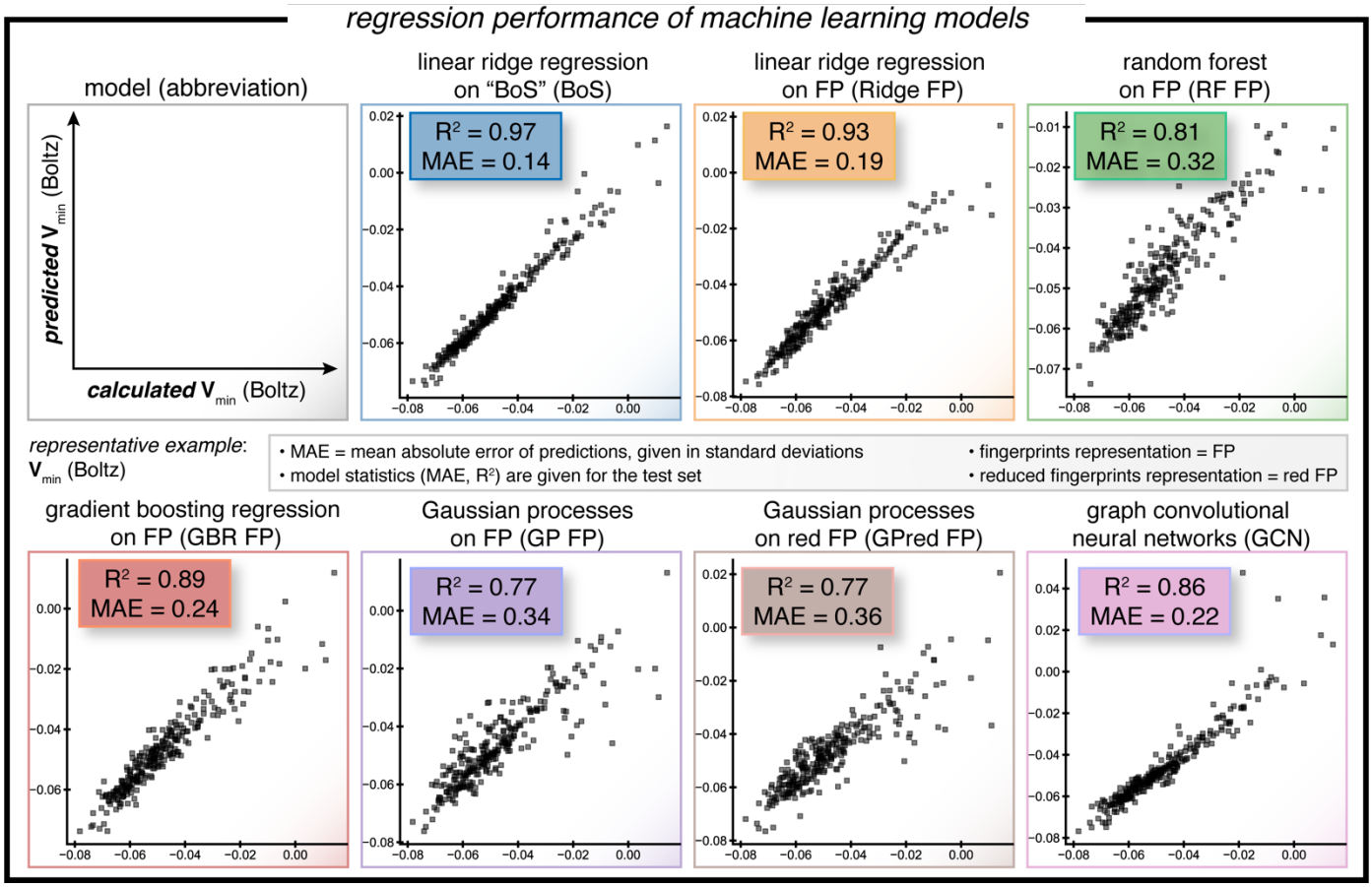

Figure 5: Regression performance of machine learning models. Illustrative performance of all seven types of ML models from this study for the prediction of V_{min}. BoS = Bag of Substituents, FP = fingerprint representation: circular fingerprints, radius = 2, folded to 1024 dimensions, red FP = reduced fingerprints representation: 100 most important fingerprint dimensions based on the feature importance of the GBR FP model. For additional details on the ML models see the ESI.

The performance of the metamodel predictors is illustrated in **Figure 6A** with V_{min} as an example. We then applied the metamodels to the >300,000 compounds arising from binary combinations of all unique substituents present in our original library (VL1) to create an extensive virtual library (VL2) with estimated descriptors. This chemical space can be visualized in a new PCA plot revealing the virtual space now available (see **Figure 6D**). Compared to the PCA plot of VL1 (cf. **Figure 3B**), the plot of VL2 appears more continuous in the descriptor ranges covered and extrapolates considerably into underexplored chemical space thereby encompassing many new structures that one might want to explore in future catalytic reactions in a single lookup table.

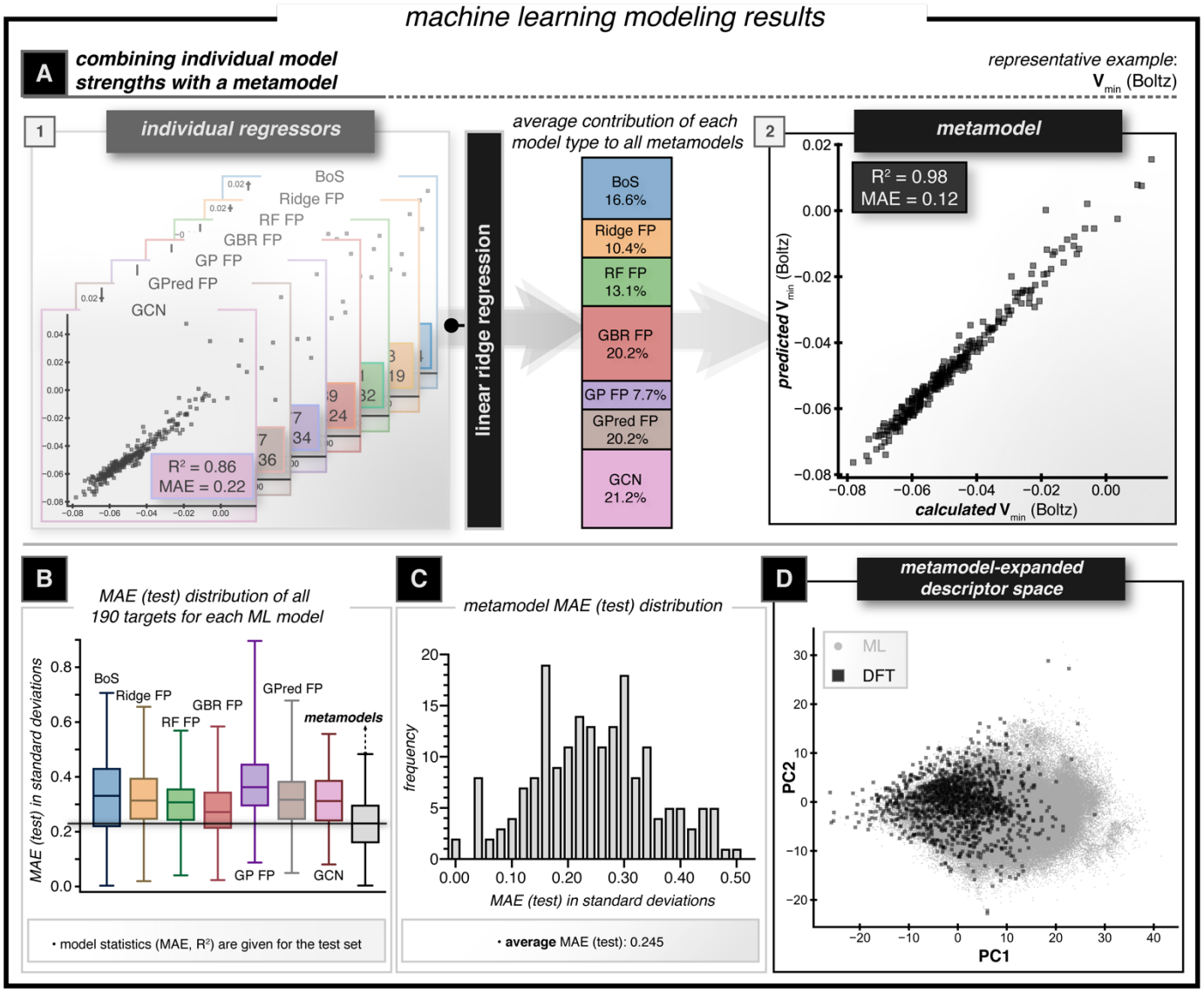

Figure 6: **Machine learning modeling results**. **A**. Stacked linear ridge regression of the seven models was used to create a metamodel for each descriptor. **B**. Comparison of the mean absolute errors (MAEs) of the seven initial model classes and the metamodels across all descriptors. **C**. Distribution of the MAEs of the metamodels across all descriptors. **D**. Expansion of the descriptor space from VL1 to VL2 with the metamodels as illustrated by PCA with VL2 being projected onto the first two principal components obtained from VL1.

Inverse Ligand Design

Finally, we aimed to demonstrate the immediate practical applicability of *kraken* to a typical problem common in reaction development and ligand design. Specifically, we wanted to utilize the ML-predicted database to identify viable alternative ligands for a selective catalytic reaction. To do this, we revisited two independent studies by Biscoe and Burke, respectively,[42,43] that reported enantiospecific Pd-catalyzed sp^3-sp^2 cross-coupling reactions of stereodefined alkyl-boronic acid derivatives with aryl halides. The two studies identified unique ligands that successfully achieve high levels of stereoretention (**Figure 7A**). In the Biscoe study, the ligand discovery was guided using predictions from statistical modeling that electron-poor Buchwald-type[44,45] biaryl phosphine ligands were the best performers. The Burke study also discovered that electron-poor ligands were required but a different core structure, one

based on *ortho*-tolyl phosphines, was found to be highly selective for this reaction. Intuitively, the best ligands from either study are structurally unique and a practicing organic chemist would not necessarily think to substitute one with the other. In addition, the reaction conditions, while distinct, are similar enough to expect qualitatively comparable selectivity of the ligands under each condition.

Based on these findings, we hypothesized that *kraken*'s descriptors applied to an original dataset could be used to predict similar ligand structures found to be optimal in the others. As a first step, several statistical models of each data set were constructed (details in the ESI) by correlating experimental results to the ligand descriptors, which were included in VL1 (**Figure 7B**). Unique models were averaged to provide robust predictions of which ligands would provide high selectivity.[46] Gratifyingly, trained on the results reported by the Burke group, our predictions identify the exact ligands reported by the Biscoe group as most selective. Similarly, regressing the Biscoe dataset and virtually screening VL1 revealed untested electron-poor triaryl phosphines, in particular, Buchwald and *ortho*-tolyl derivatives, as most selective. This suggests that these two reactions likely proceed via similar mechanistic pathways in the stereodetermining events.

After this successful validation of the interconnectivity of the two reactions, we combined the two datasets to enhance the robustness of the predictions while exploring the entire virtual search space of VL2. This is visualized in the PCA plot depicted in **Figure 7C** wherein the black-framed points represent the experimental data from the two studies, atop the ML library in gray. We were then interested in comparing two distinct approaches to suggest novel ligands in a large search space. First, we applied the averaged regression models that were trained on the experimental results to the entire VL2 to obtain selectivity predictions and robustness estimates. The ligand predictions were then curated by filtering structures through descriptor limits reported for this process (small ligands)[42] and ligands that presumably would not form a metal complex (very large ligands). As a result, we obtained ~100 ligands that are predicted to provide selective stereoretentive cross-coupling. Many of these are bulky and electron-poor Buchwald-type ligands, represented by the two structures in **Figure 7D**. Notably, ligand **D1** merges structural elements into a hybrid of both the Biscoe and Burke ligand designs.

While this approach likely provides relatively safe predictions with structural similarity to the best experimental ligands, we envisioned an explorative strategy providing more structural diversity by analyzing the relative positions of ligands in the descriptor space. We classified the "more selective" and "less selective" regions in this space defined proximity to the nearest experimental data point in the first four principle components and ranked the resulting >30,000 structures by minimizing the distance to the most selective experimental ligands. This explorative classification method suggests unexplored ligands that upon inspection have some structural familiarity to both Burke's and Biscoe's ligand designs, which is highly encouraging. This strategy would be especially effective when a researcher has relatively sparse data early in an optimization campaign as the local neighborhoods of the active space could be rapidly explored. We also envision this process will be highly effective in iterative ligand searches, especially when commercial ligands only provide modest performance.

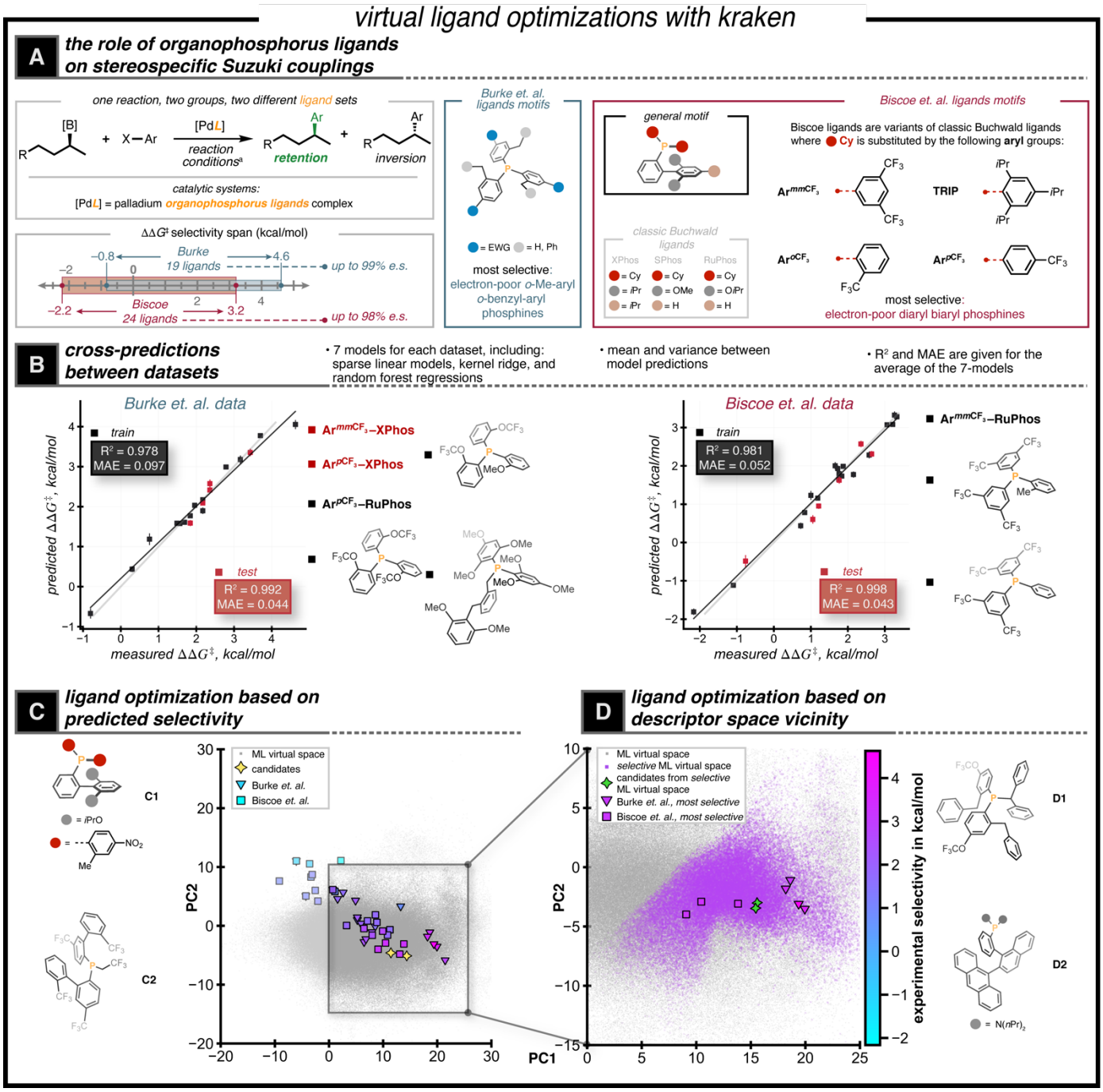

Figure 7: Using *kraken* for virtual ligand optimizations in asymmetric catalysis. A. Overview of enantiospecific Pd-catalyzed sp^3-sp^2 cross-coupling reactions of alkyl boronic acids and aryl halides. [a]: conditions: Burke *et al.*: [B] = $B(OH)_2$ (*S*), R = H, X–Ar = 1-Br-4-Ph-C_6H_4, [PdL] = Pd_2dba_3 (5 mol%) + 10 mol% L, base = Ag_2O (3 eq), solvent = dioxane, *T* = 85 °C, *t* = 24 h. Biscoe *et al.*: [B] = BF_3K (*R*), R = Ph, X–Ar = 1-Cl-4-CO_2Et-C_6H_4, [PdL] = G3 Buchwald precatalyst (10 mol%), base = K_2CO_3 (3 eq), solvent = toluene:H_2O (2:1), *T* = 100 °C, *t* = 24 h. **B**. Statistical modeling of experimental results to predict how data from one reported reaction could inform ligand choice in the other through a virtual screen of VL1. **C**. Combining the results of the statistical models for both reactions to evaluate the entirety of VL2 for predicting new ligands. **D**. Exploring the PCA descriptor space to determine ligands in the high selectivity regime with novel structures.

Conclusions and Outlook

We have developed *kraken*, which covers 300 thousand monodentate organophosphorus(III) ligands with 190 property descriptors including an extensive description of their conformer-dependence, mapping essentially the complete space of conceivable structures that could be used in organo(transition)metal reactions. We demonstrate its application in visualizing the associated property space, predicting properties of molecules not subjected to our full quantum-chemical workflow, and applying the corresponding results to inverse catalyst design.

Kraken is accessible as a web application (https://kraken.cs.toronto.edu). Computed data is available at the semi-empirical QM, DFT, and ML levels of theory. For 1,558 organophosphorus compounds, there are both semi-empirical QM and DFT data comprising 190 computed descriptors and properties, as well as the coordinates information for the associated conformers. The ML data consists of 331,776 entries obtained by generating all organophosphorus ligands with two distinct substituents combinatorially and training the models on the Boltzmann-averaged DFT dataset (see above). Lastly, around 183 million distinct organophosphorus compounds can be queried to generate the ML property predictions on-the-fly.

Overall, we believe that the property maps generated by common dimensionality reduction techniques included in the *kraken* platform can be a valuable aid in the understanding of the space of organophosphorus ligands. We envision that it will enable organic chemists to perform computer-assisted interactive ligand exploration and provide new insights into relevant properties to solve a given problem. The *kraken* tool may enable informed catalyst design based on organophosphorus ligands, facilitate the optimization of reaction process parameters, inspire new ligand choices and promote the synthesis of new organophosphorus compounds. The database and tools reported herein are currently being applied to enhance reaction optimization[47] and mechanistic workflows.[48] The open-source nature of our codes, as well as the open database, is designed to be extended by others and we welcome further contributions by the community.

Acknowledgments

T.Ge. thanks the Leopoldina Fellowship Programme of the German National Academy of Sciences Leopoldina (LPDS 2017−18). T.Ge. is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2008/1 – 390540038. T.Ge. wird durch die Deutsche Forschungsgemeinschaft (DFG) im Rahmen der Exzellenzstrategie des Bundes und der Länder – EXC 2008/1 – 390540038 gefördert. G.P.G gratefully acknowledges the Natural Sciences and Engineering Research Council of Canada (NSERC) for the Banting Postdoctoral Fellowship. R.P. acknowledges funding through a Postdoc.Mobility fellowship by the Swiss National Science Foundation (SNSF, Project No. 191127). K.J. was a fellow of the AstraZeneca Postdoc Programme (2018–2020). M.L.D. gratefully acknowledges the Fonds de Recherche Quebec Nature et Technologies (FRQNT) for the B1X Master's Scholarship and support from the Queen Elizabeth II Graduate Scholarship in Science and Technology (QEII-GSST). The support and resources from the Center for High Performance Computing at the University of Utah are gratefully acknowledged.

We acknowledge the Defense Advanced Research Projects Agency (DARPA) under the Accelerated Molecular Discovery Program under Cooperative Agreement No. HR00111920027 dated August 1, 2019. The content of the information presented in this work does not necessarily reflect the position or the policy of the Government. A.A.-G. thanks Anders G. Frøseth for his generous support. A.A.-

G. also acknowledges the generous support of Natural Resources Canada and the Canada 150 Research Chairs program.

We thank *Compute Canada* for computational resources. DFT and *xtb* calculations were performed on the *niagara* supercomputer at the SciNet HPC Consortium. SciNet is funded by: the Canada Foundation for Innovation; the Government of Ontario; Ontario Research Fund - Research Excellence; and the University of Toronto. Machine learning models were developed and trained on the supercomputer *beluga* from École de technologie supérieure, managed by Calcul Québec and Compute Canada. The operation of this supercomputer is funded by the Canada Foundation for Innovation (CFI), the ministère de l'Économie, de la science et de l'innovation du Québec (MESI) and the Fonds de recherche du Québec – Nature et technologies (FRQ-NT).

We are grateful to UofT Matter Lab system administrators Dr. Claire Yu and Chris Crebolder for helping with the deployment of the web app.

M.S.S. and E.P. thank the NSF under the CCI Center for Computer Assisted Synthesis (CHE-1925607) for support.

Author Contributions

T.Ge., G.P.G., P.F., K.J., M.S.S., and A.A.-G. conceptualized this project. T.Ge., G.P.G., and P.F. built the workflows, performed the majority of the quantum-chemical calculations, and developed the machine learning models. E.P. contributed to several discussions related to this project, solidified the workflows, and organized the ESI. T.Ga. is the main developer of the web app and contributed to several discussions of this work. R.P. contributed to the development of the property calculation workflow and to several discussions related to this project. K.J. is the main developer of MORFEUS and contributed to several discussions of this work. A.N. assisted with the development of the graph-based machine learning models and contributed to several discussions related to this project. M.L.D. performed quantum-chemical calculations and solidified the codes used in this work. All authors contributed to the writing of this manuscript.

Web Application and Codes Availability

The database can be accessed and used free of charge via the web application at https://kraken.cs.toronto.edu. MORFEUS can be freely accessed at https://github.com/kjelljorner/morfeus. The collection of workflow codes and machine learning models used in this project will be available at https://github.com/aspuru-guzik-group/kraken.

Conflicts of interest

A.A.-G. is a co-founder and the Chief Visionary Officer at Kebotix Inc.

References and Notes

(1) Milo, A.; Neel, A. J.; Toste, F. D.; Sigman, M. S. *Science* **2015**, *347*, 737–743.
(2) Sanchez-Lengeling, B.; Aspuru-Guzik, A. *Science* **2018**, *361*, 360–365.
(3) David, L.; Thakkar, A.; Mercado, R.; Engkvist, O. *Journal of Cheminformatics* **2020**, *12*, 56.
(4) Goscinski, A.; Fraux, G.; Imbalzano, G.; Ceriotti, M. *Mach. Learn.: Sci. Technol.* **2021**, *2*, 025028.
(5) Ahneman, D. T.; Estrada, J. G.; Lin, S.; Dreher, S. D.; Doyle, A. G. *Science* **2018**, *360*, 186–190.
(6) Zahrt, A. F.; Henle, J. J.; Rose, B. T.; Wang, Y.; Darrow, W. T.; Denmark, S. E. *Science* **2019**, *363*.
(7) John, P. C. S.; Guan, Y.; Kim, Y.; Kim, S.; Paton, R. S. *Nature Communications* **2020**, *11*, 2328.
(8) Santiago, C. B.; Guo, J.-Y.; Sigman, M. S. *Chem. Sci.* **2018**, *9*, 2398–2412.
(9) Wen, M.; Blau, S. M.; Spotte-Smith, E. W. C.; Dwaraknath, S.; Persson, K. A. *Chem. Sci.* **2021**, *12*, 1858–1868.
(10) Pyzer-Knapp, E. O.; Li, K.; Aspuru-Guzik, A. *Advanced Functional Materials* **2015**, *25*, 6495–6502.
(11) Jain, A.; Ong, S. P.; Hautier, G.; Chen, W.; Richards, W. D.; Dacek, S.; Cholia, S.; Gunter, D.; Skinner, D.; Ceder, G.; et al. *APL Materials* **2013**, *1*, 011002.

(12) Kirklin, S.; Saal, J. E.; Meredig, B.; Thompson, A.; Doak, J. W.; Aykol, M.; Rühl, S.; Wolverton, C. *npj Computational Materials* **2015**, *1*, 1–15.
(13) Curtarolo, S.; Setyawan, W.; Hart, G. L. W.; Jahnatek, M.; Chepulskii, R. V.; Taylor, R. H.; Wang, S.; Xue, J.; Yang, K.; Levy, O.; et al. *Computational Materials Science* **2012**, *58*, 218–226.
(14) Hachmann, J.; Olivares-Amaya, R.; Atahan-Evrenk, S.; Amador-Bedolla, C.; Sánchez-Carrera, R. S.; Gold-Parker, A.; Vogt, L.; Brockway, A. M.; Aspuru-Guzik, A. *J. Phys. Chem. Lett.* **2011**, *2*, 2241–2251.
(15) Winther, K. T.; Hoffmann, M. J.; Boes, J. R.; Mamun, O.; Bajdich, M.; Bligaard, T. *Scientific Data* **2019**, *6*, 75.
(16) Chanussot, L.; Das, A.; Goyal, S.; Lavril, T.; Shuaibi, M.; Riviere, M.; Tran, K.; Heras-Domingo, J.; Ho, C.; Hu, W.; et al. *arXiv:2010.09990 [cond-mat]* **2021**.
(17) Tolman, C. A. *J. Am. Chem. Soc.* **1970**, *92*, 2953–2956.
(18) Tolman, C. A. *J. Am. Chem. Soc.* **1970**, *92*, 2956–2965.
(19) Hansch, Corwin.; Leo, A.; Taft, R. W. *Chem. Rev.* **1991**, *91*, 165–195.
(20) Benson, S. W.; Cruickshank, F. R.; Golden, D. M.; Haugen, G. R.; O'Neal, H. E.; Rodgers, A. S.; Shaw, R.; Walsh, R. *Chem. Rev.* **1969**, *69*, 279–324.
(21) Fey, N.; Tsipis, A. C.; Harris, S. E.; Harvey, J. N.; Orpen, A. G.; Mansson, R. A. *Chemistry – A European Journal* **2006**, *12*, 291–302.
(22) Jover, J.; Fey, N.; Harvey, J. N.; Lloyd-Jones, G. C.; Orpen, A. G.; Owen-Smith, G. J. J.; Murray, P.; Hose, D. R. J.; Osborne, R.; Purdie, M. *Organometallics* **2010**, *29*, 6245–6258.
(23) Durand, D. J.; Fey, N. *Chem. Rev.* **2019**, *119*, 6561–6594.
(24) Crawford, J. M.; Sigman, M. S. *Synthesis* **2019**, *51*, 1021–1036.
(25) Brethomé, A. V.; Fletcher, S. P.; Paton, R. S. *ACS Catal.* **2019**, *9*, 2313–2323.
(26) Fey, N.; Harvey, J. N.; Lloyd-Jones, G. C.; Murray, P.; Orpen, A. G.; Osborne, R.; Purdie, M. *Organometallics* **2008**, *27*, 1372–1383.
(27) Wilbraham, L.; Berardo, E.; Turcani, L.; Jelfs, K. E.; Zwijnenburg, M. A. *J. Chem. Inf. Model.* **2018**, *58*, 2450–2459.
(28) Hillier, A. C.; Sommer, W. J.; Yong, B. S.; Petersen, J. L.; Cavallo, L.; Nolan, S. P. *Organometallics* **2003**, *22*, 4322–4326.
(29) Grimme, S.; Bannwarth, C.; Shushkov, P. *J. Chem. Theory Comput.* **2017**, *13*, 1989–2009.
(30) Bannwarth, C.; Ehlert, S.; Grimme, S. *J. Chem. Theory Comput.* **2019**, *15*, 1652–1671.
(31) Grimme, S. *J. Chem. Theory Comput.* **2019**, *15*, 2847–2862.
(32) Pracht, P.; Bohle, F.; Grimme, S. *Phys. Chem. Chem. Phys.* **2020**, *22*, 7169–7192.
(33) Tolman, C. A. *Chem. Rev.* **1977**, *77*, 313–348.
(34) McInnes, L.; Healy, J.; Melville, J. *arXiv:1802.03426 [cs, stat]* **2020**.
(35) Suresh, C. H.; Koga, N. *Inorg. Chem.* **2002**, *41*, 1573–1578.
(36) Diorazio, L. J.; Hose, D. R. J.; Adlington, N. K. *Org. Process Res. Dev.* **2016**, *20*, 760–773.
(37) Benson, S. W.; Buss, J. H. *J. Chem. Phys.* **1958**, *29*, 546–572.
(38) Friederich, P.; Krenn, M.; Tamblyn, I.; Aspuru-Guzik, A. *Mach. Learn.: Sci. Technol.* **2021**, *2*, 025027.
(39) Friederich, P.; Gomes, G. dos P.; Bin, R. D.; Aspuru-Guzik, A.; Balcells, D. *Chem. Sci.* **2020**, *11*, 4584–4601.
(40) Lopez, S. A.; Sanchez-Lengeling, B.; de Goes Soares, J.; Aspuru-Guzik, A. *Joule* **2017**, *1*, 857–870.
(41) Duvenaud, D. K.; Maclaurin, D.; Iparraguirre, J.; Bombarell, R.; Hirzel, T.; Aspuru-Guzik, A.; Adams, R. P. In *Advances in Neural Information Processing Systems 28*; Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., Garnett, R., Eds.; Curran Associates, Inc., 2015; pp 2224–2232.
(42) Zhao, S.; Gensch, T.; Murray, B.; Niemeyer, Z. L.; Sigman, M. S.; Biscoe, M. R. *Science* **2018**, *362*, 670–674.
(43) Lehmann, J. W.; Crouch, I. T.; Blair, D. J.; Trobe, M.; Wang, P.; Li, J.; Burke, M. D. *Nature Communications* **2019**, *10*, 1263.
(44) Surry, D. S.; Buchwald, S. L. *Chem. Sci.* **2010**, *2*, 27–50.
(45) Ingoglia, B. T.; Wagen, C. C.; Buchwald, S. L. *Tetrahedron* **2019**, *75*, 4199–4211.
(46) Brethomé, A. V.; Paton, R. S.; Fletcher, S. P. *ACS Catal.* **2019**, *9*, 7179–7187.
(47) Christensen, M.; Yunker, L.; Adedeji, F.; Häse, F.; Roch, L.; Gensch, T.; dos Passos Gomes, G.; Zepel, T.; Sigman, M.; Aspuru-Guzik, A.; et al. *ChemRxiv* **2020**, 10.26434/chemrxiv.13146404.v2.
(48) Newman-Stonebraker, S. H.; Smith, S. R.; Borowski, J. E.; Gensch, T.; Peters, E. B.; Johnson, H. C.; Sigman, M. S.; Doyle, A. G. *ChemRxiv* **2021**, 10.26434/chemrxiv.14388557.