Chemical Science

ROYAL SOCIETY OF CHEMISTRY

EDGE ARTICLE

View Article Online
View Journal | View Issue



Cite this: *Chem. Sci.*, 2022, **13**, 13782

All publication charges for this article have been paid for by the Royal Society of Chemistry

Received 29th July 2022
Accepted 24th October 2022

DOI: 10.1039/d2sc04251g

rsc.li/chemical-science



# OSCAR: an extensive repository of chemically and functionally diverse organocatalysts†

Simone Gallarati,[a] Puck van Gerwen,[ab] Ruben Laplaza,[ab] Sergi Vela,[a] Alberto Fabrizio[ac] and Clemence Corminboeuf *[abc]

The automated construction of datasets has become increasingly relevant in computational chemistry. While transition-metal catalysis has greatly benefitted from bottom-up or top-down strategies for the curation of organometallic complexes libraries, the field of organocatalysis is mostly dominated by case-by-case studies, with a lack of transferable data-driven tools that facilitate both the exploration of a wider range of catalyst space and the optimization of reaction properties. For these reasons, we introduce OSCAR, a repository of 4000 experimentally derived organocatalysts along with their corresponding building blocks and combinatorially enriched structures. We outline the fragment-based approach used for database generation and showcase the chemical diversity, in terms of functions and molecular properties, covered in OSCAR. The structures and corresponding stereoelectronic properties are publicly available (https://archive.materialscloud.org/record/2022.106) and constitute the starting point to build generative and predictive models for organocatalyst performance.

## Introduction

Constructing extensive yet tailored databases is crucial for the successful development and application of data-driven tools in catalysis and materials science.[1,2] The way datasets are generated largely reflects how chemists think about the structure of a catalyst. In turn, this not only influences the way improved molecular systems are searched, but also how their structure is manipulated, for example through trial-and-error,[3] fine-tuning according to mechanistic insight,[4–7] or generating compound libraries for activity/selectivity screening.[8–11]

Transition-metal catalysts are naturally viewed in a modular fashion as a combination of active metal centre and ligands, which are further decomposed into metal-coordinating groups, backbone/bridging units, and substituents.[12] This simple, yet powerful fragment-based strategy has enabled tremendous advancements in computer-aided catalyst design,[13,14] from the exploration of the chemical space of inorganic species curated through bottom-up or top-down approaches,[15–20] the construction of ligand databases with associated steric and electronic descriptors,[21–27] to the development of algorithms for the assembly of metal complexes from fragments and evolutionary experiments.[28–30] Modularity is even more apparent in biocatalysts,[31] which combine a limited number of building blocks, the amino acids; inspired by natural evolution, strategies such as combinatorial backbone assembly[32] have allowed to generate libraries of structurally diverse enzymes with altered catalytic properties.

Organocatalysts are far less frequently classified according to fragment-based schemes. Instead, they are typically grouped into families of "privileged catalysts",[33,34] or according to the functional components that encapsulate their catalytic power (Fig. 1).[35] Privileged catalysts are those species possessing certain chiral scaffolds that have proven to be effective at inducing high levels of enantioselectivity across a wide range of mechanistically unrelated reactions.[33,34] Some effort has been made to summarize these catalytic motifs,[36] however their comprehensive enumeration across all of chemical space is challenging due to the large possible variations in functionalities. This problem is exacerbated by the fact that organocatalysts are essentially a subclass of organic molecules, whose space is estimated to exceed $10^{60}$,[37,38] and chemical expertise is required to evaluate whether an organic molecule could function as a catalyst in a reaction. Therefore, *de novo* organocatalyst design is a formidable, seldomly approached task, primarily due to the lack of robust ways of defining and assembling their building blocks,[39,40] and reaction optimization is dominated by testing closely related analogues of a known privileged catalyst.[41]

[a]*Laboratory for Computational Molecular Design, Institute of Chemical Sciences and Engineering, Ecole Polytechnique Fédérale de Lausanne (EPFL), 1015 Lausanne, Switzerland. E-mail: clemence.corminboeuf@epfl.ch*

[b]*National Center for Competence in Research – Catalysis (NCCR-Catalysis), Ecole Polytechnique Fédérale de Lausanne (EPFL), 1015 Lausanne, Switzerland*

[c]*National Center for Computational Design and Discovery of Novel Materials (MARVEL), Ecole Polytechnique Fédérale de Lausanne (EPFL), 1015 Lausanne, Switzerland*

† Electronic supplementary information (ESI) available. See DOI: https://doi.org/10.1039/d2sc04251g

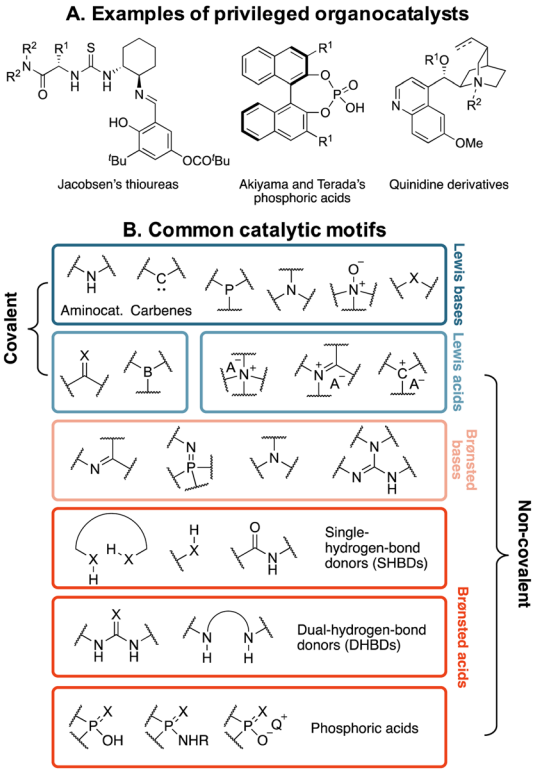


Fig. 1 (A) Prototypical privileged chiral frameworks for asymmetric catalysis. (B) Classification of organocatalysts according to their catalytic motifs (X = O, S).

Similarly, the field of data-driven organocatalysis has been dominated by efforts, either on the automation side[42] (*e.g.*, AARON[43] or ACE/Virtual Chemist[44,45]) or on the development of statistical models for enantioselectivity prediction,[46–52] that have focused on specific reaction classes or structurally related catalysts. There is currently a dearth of general strategies and platforms for organocatalysts comparison, fragmentation into building blocks, and assembly across a wide region of catalyst space, encompassing functionally and chemically diverse molecules with a multitude of catalytic functions.

In this work, we propose a solution in the form of OSCAR (Organic Structures for CAtalysis Repository), a database of experimentally derived or combinatorially enriched organocatalysts and of the corresponding molecular fragments that are extracted from them. Not only OSCAR constitutes a map to navigate organocatalyst space and potentially enable informed catalyst design, but the modular strategy behind its construction paves the way to a multitude of data-driven and fragment-based reaction optimization methods.[53,54] Herein, we show how such a dataset is curated and augmented with crystallographically determined structures using a combination of top-down and bottom-up approaches, and how the fragments are assembled in a combinatorial fashion to generate thousands of species. In its current forms, OSCAR contains 4000 catalysts, whose use has either been documented in the literature for organic synthesis or with chemically analogous structure reported in the Cambridge Structural Database (CSD), spanning various catalytic functions (Lewis/Brønsted acids and bases), and two exemplary enriched combinatorial supersets, OSCAR!(NHC) and OSCAR!(DHBD). The former consists of over 8000 carbenes for covalent catalysis, the latter contains *ca.* 1.5 million non-covalent dual-hydrogen-bond donors. The approaches used to generate these combinatorial databases (*vide infra*) are however transferable to other classes, implying the possibility of further extending OSCAR. A selection of stereoelectronic molecular descriptors, including reactivity indices derived from conceptual DFT, are provided and may help establishing structure–reactivity relationships for reaction optimization. All structures and properties are publicly available on the Materials Cloud for interactive visualization with Chemiscope (**https://doi.org/10.24435/materialscloud:gy-3h**).[55] They could serve as the starting point to define the combinatorial space for evolutionary experiments,[56] as well as the basis for dataset curation to train machine learning models for applications in organic synthesis.[54]

## Results and discussion

### Database curation

No comprehensive repository of organocatalysts' structures covering all of the functionalities summarized in Fig. 1B currently exists. Most frequently, they are reported in the literature in 2D format (*i.e.*, ChemDraw pictures) with associated experimental characterization data in the ESI† (NMR and IR spectra and, less often, crystal structure information), but molecular geometries are not easily accessible. To construct OSCAR, we followed a five-step protocol (Fig. 2), which starts with the manual collection of catalysts (as 2D objects) from reviews,[35,57–68] journal articles,[69–72] books,[73–77] and commercial

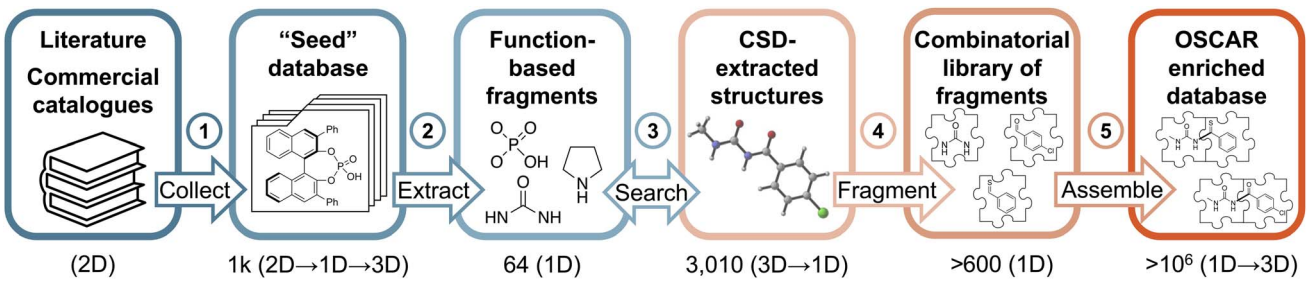


Fig. 2 Graphical summary of the steps followed for the curation of OSCAR.

catalogues[78,79] into a "seed" database (step 1). Each of the 1000 2D entries in this library is labelled according to the classes in Fig. 1 and converted into a 1D (*i.e.*, SMILES strings) and subsequently 3D (*i.e.*, optimized XYZ geometry) structure (see the Computational methods). Given that more than ~1500 publications on organocatalysis are published each year,[80] it is virtually impossible to curate an exhaustive library of all existing catalysts. Nonetheless, the seed database aims at covering the chemical diversity observed across all of organocatalyst space in terms of chemical functionalities, catalytic motifs and scaffolds/substituents, with the added bonus of each structure either being commercially available or synthetically accessible, having being mined from the literature.

This top-down approach ensures that only organic molecules that have been reported to display, or be tested for, catalytic activity are included in the database. However, it is a slow, human error-prone process that cannot be automated and might either introduce in the repository erroneous or mis-labelled structures or lead to chemically interesting ones being excluded. Existing crystallographic databases (*e.g.*, CSD,[81,82] COD[83]) offer the most comprehensive collection of organic (and inorganic) molecules that have been synthetised. Although it not possible to filter out *a priori* those compounds that have not been tested as organocatalysts, CSD offers the chance to significantly augment the seed database with more, chemically diverse structures, provided that the right chemical motifs, which might make a molecule catalytically active, are searched. To achieve this goal we enumerated, in 1D format, 64 "function-based fragments" included in the seed database (step 2 Fig. 2 and S1 and S2†). Although not exhaustive, they represent the most common catalytic motifs and ensure that the species that contain them are relevant to the task at hand. In step 3, these fragments are searched in CSD and the corresponding whole molecules extracted. After retrieving the 3D geometries from the cif files with the cell2mol software,[84] 3010 compounds are added to the seed database, yielding a total of 4000 entries (after filtering out identical ones, see the ESI†). All 3D entries are then converted into 1D format for subsequent fragmentation and recombination (steps 4 and 5, *vide infra*).

With respect to the catalytic motifs (*cf.* Fig. 1B), the distribution of the CSD-extracted structures changes significantly from the seed database (see the two histograms in Fig. 3A). In OSCAR, the majority of species (40%) are classified as dual-hydrogen-bond donors; their large increase in number upon CSD extraction is likely due to the popularity of the (thio)urea moiety as pharmacophore[85–87] and for anion recognition.[88] The second most popular class (24%) is aminocatalysts based on the pyrrolidine motif: in the early days of organocatalysis, the vast majority of reactions were indeed amine-based[59,89] and five-membered (polycyclic) secondary amines are widely encountered in natural products, as well as being a preferred scaffold in pharmaceutical science and drug design.[90] The other classes are more or less equally represented (~5–6%, Fig. 3B), with a slight predominance of Lewis bases (11%), given the large variety of N(O)-, P(O)-, and S(O)-nucleophilic organocatalysts. If we consider the increase in type of heteroatoms from the seed to the CSD-extracted database (Fig. 3C), sulphur and nitrogen are

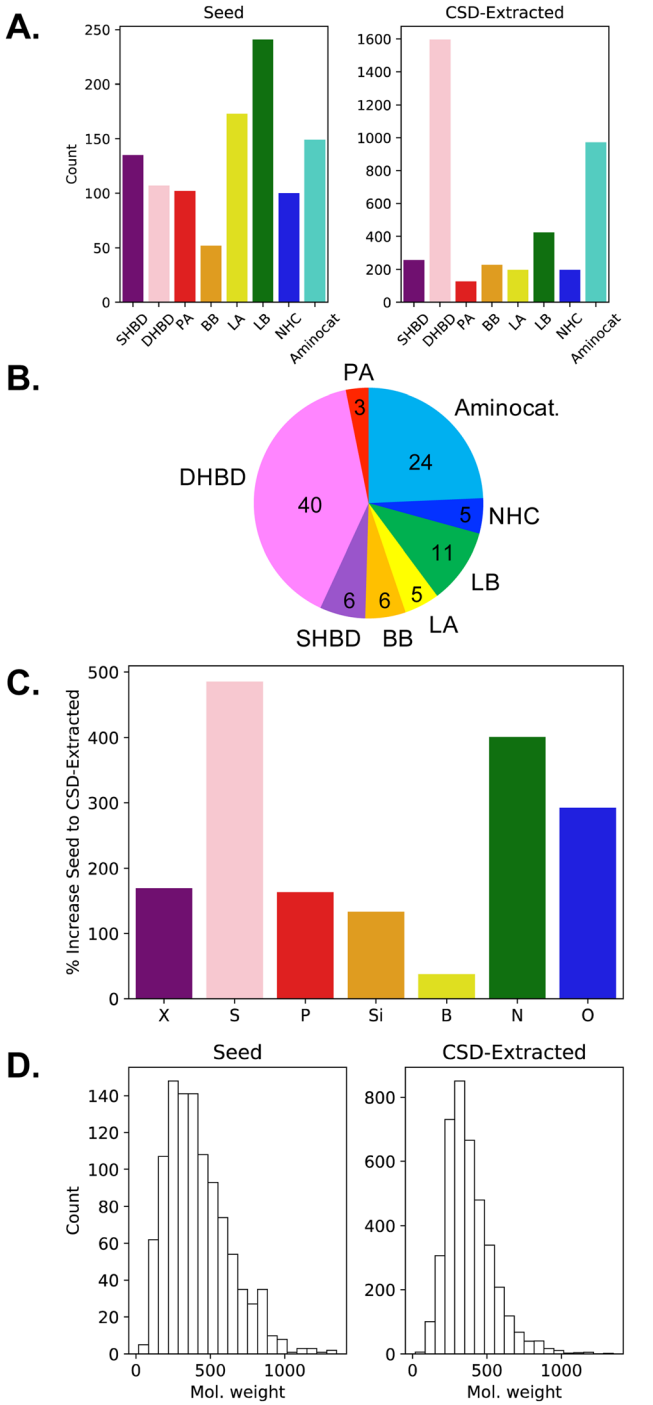


Fig. 3 (A) Distribution histograms of catalytic motifs in the seed database and in the CSD-extracted structures. (B) Pie chart showing percentages of catalytic motifs in the seed and CSD-extracted datasets. (C) Distribution histograms of heteroatom types (X = halogens), and (D) molecular weight in the seed and in the CSD-extracted sets.

the most abundant due to the predominance of the thiourea and pyrrolidine catalytic motifs. The amount of P, Si, X, and especially B atoms increases to a significantly lesser extent. In the case of phosphorous, even though we seek to augment the quantity of P-containing motifs, only a limited number of phosphoric acids (*ca.* 25) are extractable from CSD. On the other

hand, no catalytic unit that specifically contains halogens, silicon or boron is searched. An exhaustive description of the functional groups present in OSCAR is given in the ESI (Table S2 and Fig. S4†). Finally, the catalysts in the two datasets have a similar distribution of molecular weights (Fig. 3D), with the seed database containing on average slightly larger molecules (~430 u) and a displaying smoother decrease in occurrences as their size increases.

### Structure and property maps

The chemical and structural diversity contained in OSCAR is visualized in Fig. 4A with a 2D t-SNE map[91] based on FCHL19[92]

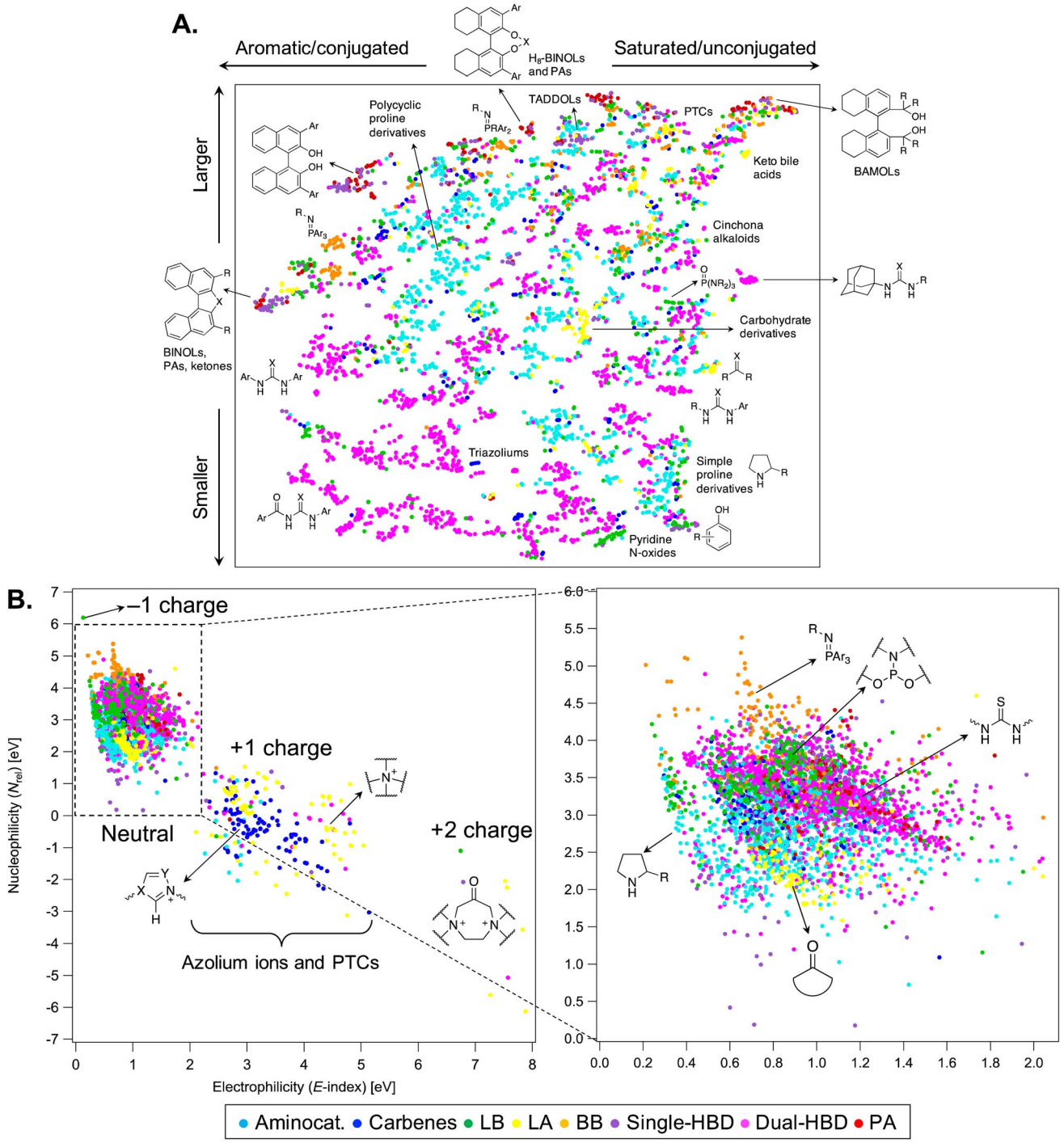


Fig. 4 (A) 2D t-SNE map of OSCAR on the basis of the FCHL19 representation.[92] Each point represents an organocatalyst, coloured by the corresponding catalytic motif. Each cluster contains catalysts with similar structure, with some examples being shown. R = alkyl group; Ar = aromatic group; PTC = phase-transfer catalyst. (B) Property map: computed (ωB97X-D/Def2-TZVP//B97-D/Def2-TZVP) nucleophilicity ($N_{rel}$) *vs.* electrophilicity (*E*-index) parameters.[98] A zoom-in of the map is provided on the right hand side.

of the 4000 organocatalysts from the seed and CSD databases. Alternative representations and dimensionality reduction can be found in the ESI (Fig. S5–S7†). Although the two axes (dimensions) of this structure map have no formal physical meaning, it is possible to establish a qualitative relationship between them and chemical properties. In particular, species found higher in the map are bigger (higher molecular weight/surface area), whereas the degree of conjugation and the presence of aromatic scaffolds and substituents decreases left to right. For example, diol-based catalysts,[93] which act as single-hydrogen-bond donors, and phosphoric acids[94] are found along the upper edge of the map, with the fully aromatic BINOL derivatives on the left, the $H_8$-BINOL core in the middle, and BAMOLs on the right. Dual-HBDs, especially diaryl (thio)ureas, occupy the lower left corner of the map, while simple proline derivatives the bottom right, with larger and more complex aminocatalysts in the upper left region. Other noticeable clusters correspond to the ketone epoxidation catalysts developed by Shi and Shu (covalent Lewis acidic carbohydrate derivatives),[95,96] and to iminophosphorane Brønsted bases.[97]

The structure map is complemented by a "property map" (Fig. 4B) in which the organocatalysts are evaluated in terms of their DFT-computed global electro/nucleophilicity indices (see the Computational methods), which assume that, when these catalysts react, they do so cumulatively and simultaneously at all their atomic sites.[99] The largest influence on the descriptors is exerted by the total molecular charge, and three regions are found (four if the green point corresponding to the phosphorylated sulfonimidamide[100] with −1 charge is considered). *E*-index increases with the charge, while $N_{rel}$ decreases. Highly electrophilic and charged species include phase transfer catalysts[101] (PTCs, non-covalent Lewis acids) and azolium ions, which are the conjugate acid precursors of carbene organocatalysts.[102] The zoom-in on the right hand side of Fig. 4B shows the spread of *E*-index and $N_{rel}$ values for neutral organocatalysts. Among the most nucleophilic species, Brønsted bases, in particular iminophosphoranes, and phosphoramidite Lewis bases are found towards the top of the map ($\overline{N_{rel}} = 3.8$ eV, $\sigma = 0.6$ eV), while ketone epoxidation electrophiles are at the bottom ($\overline{E_{index}} = 1.0$ eV, $\sigma = 0.2$ eV, $\overline{N_{rel}} = 2.4$ eV, $\sigma = 0.5$ eV). Some families of catalysts, such as DHBDs containing the thiourea motif and aminocatalysts, cover a wide range of values ($0.4 < E\text{-index}_{DHBD} < 2.1$ eV), indicating that their electronic properties are highly dependent on the nature of the substituents bound to the catalytic motif. Although it is unlikely that these simple reactivity indices can accommodate a robust and universal scale for electrophilicity and nucleophilicity of such diverse molecules with a varied range of structural, electronic, and bonding properties, the property map in Fig. 4B and the set of descriptors provided with OSCAR may supplement existing structure–reactivity scales in organocatalysis,[103–109] such as the ones developed by Mayr *et al.*[110–113]

## Combinatorial databases

OSCAR currently covers a significant part of organocatalyst space and a large pool of chemically and functionally diverse catalytic motifs. However, given the nearly infinite number of possible derivatives of each catalyst, only relatively few examples are included. Harnessing the fragment-based strategy used to enrich the seed database with structures from CSD in a bottom-up fashion, we exponentially increase the size of OSCAR by building combinatorial databases from molecular fragments. The exact nature of the fragments depends on the family of organocatalysts, but they can be grouped into two categories: catalytic motifs (*i.e.*, the chemical groups that contain the reactive components) and structural substituents (which modulate their stereoelectronic properties). If the catalytic motif is easily distinguishable from the rest of the molecule (*e.g.*, for dual-hydrogen-bond donors, *vide infra*), it is extracted as a subgraph of the whole catalyst, and the rest handled as structural substituents. If the catalytic motif exhibits larger chemical diversity and substitution patterns (*e.g.*, carbenes, *vide infra*), the possible functional units and substituents are curated manually based on chemical expertise. Herein, we show how to do this for two types of covalent and non-covalent organocatalysts, specifically N-heterocyclic carbenes [OSCAR!(NHC)] and dual-hydrogen-bond donors [OSCAR!(DHBD)]. In the first case, a relatively "small" database (8622 catalysts) is curated by carefully selecting catalytic motifs and substituents found in OSCAR. In the second, we adopt a graph-based approach to generate 1 573 015 DHBDs.

In the first example (Fig. 5, top), 17 cores/scaffolds are extracted from the seed and CSD libraries (Fig. 4A and S9,† most central ring system generated with DataWarrior[114]); based on structural features reported in the literature,[115–118] 60 substituents grouped into three categories ($R^{1-3}$, Fig. S10–S12†) and appropriate substitution patterns are defined. They are then translated into flexible SMILES strings (Table S4†), written in such a way that different $R^{1-3}$ in each core can easily be introduced and exchanged. Finally, 3D structures are generated from the SMILES and fully optimized, yielding a database of 8622 species. In the second example (Fig. 5, bottom), all the organocatalysts containing one DHBD unit in the seed and CSD-extracted datasets (1593) are interpreted as molecular graphs[119] (*i.e.*, undirected multigraphs with RDKit) and fragmented into the central catalytic motif and the two substituents on either side ($R^{1,2}$), affording a combinatorial space of $7 \times 694^2$ groups. After duplicate removal and recombination with RDKit, they yield a total of 1 573 015 species (all optimized at the xTB level); 1000 structures per each DHBD motif are selected and optimized with DFT, and 6994 are shown in Fig. 6B.

The two combinatorial datasets are visualized with chemical space maps (Fig. 6),[120] which are typically constructed from steric and electronic molecular descriptors. Based on their popularity and chemical meaningfulness, the percentage of buried volume[121,122] $\%V_{buried}$ and nucleophilicity *N*-index (see the Computational methods) are the parameters chosen for OSCAR!(NHC), while the LUMO energy $\varepsilon_{LUMO}$ and the HNNH dihedral angle of the HBD unit ($\theta$) are plotted for OSCAR!(DHBD). The electronic descriptors provide an indirect estimate of the catalysts' Brønsted acidity/basicity: analysis of the experimental equilibrium acidities of 23 NHCs[123] shows that the $pK_a$ values of their precursors (the azolium ions) are directly

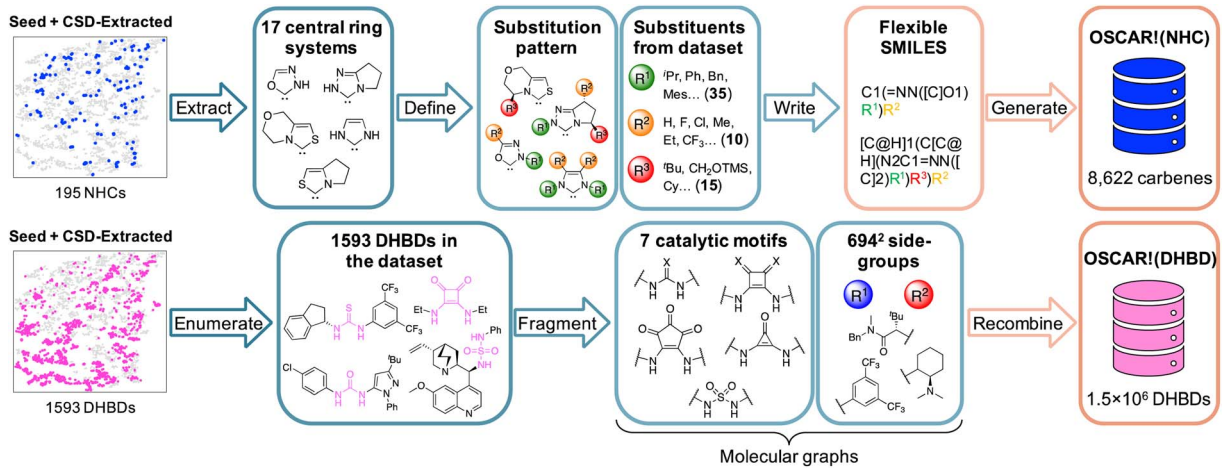


Fig. 5 Graphical summary of the steps followed to generate the combinatorial databases OSCAR!(NHC) (top) and OSCAR!(DHBD) (bottom). X = O/S.

proportional to the *N*-index of the carbene ($R^2 = 0.80$, $2\sigma = 0.72$, Fig. S8†), while the LUMO energies of 74 DHBDs[69] scale linearly ($R^2 = 0.92$, $2\sigma = 2.32$, Fig. S13†) with their experimental p$K_a$'s (as previously noted by Sigman for a smaller subset).[50] %$V_{buried}$ and $\theta$ quantify the steric influence exerted by the catalysts' core and substituents.

The NHCs in Fig. 6A are coloured according to common structural features. The N-substituent ($R^1$ in Fig. 5 and S10†) has the greatest effect on nucleophilicity, with catalysts bearing electron-donating alkyl groups [*i.e.*, Me, Et, $^i$Pr, Cy, and C(Me)Cy] having the highest *N*-index (blue points). These species are predicted to be the most reactive towards electrophilic attack, however their precursors have p$K_a$'s over 20,[123] meaning that relatively strong bases must be used for active catalyst generation. The steric demand of the carbene is mostly influenced by $R^3$ (Fig. S12†): L-pyroglutamic acid-derived bicyclic NHCs[124] with diaryl- and diaryl(hydroxy)methyl substituents[125,126] (red and purple points) are located towards the top of the map (large %

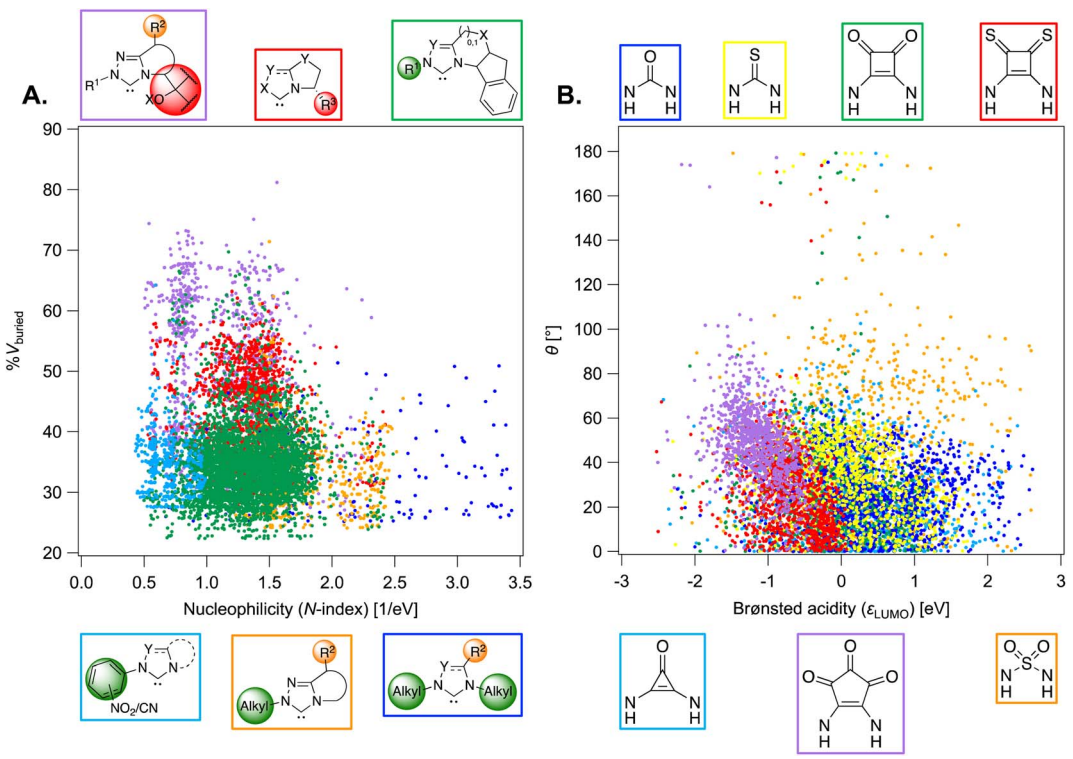


Fig. 6 (A) Percentage buried volume *vs.* *N*-index of combinatorial NHC organocatalysts. *N*-index is found to scale linearly with known experimental p$K_a$ values of azolium ions (Fig. S8†). (B) HNNH dihedral angle ($\theta$) *vs.* LUMO energy ($\omega$B97X-D/Def2-TZVP//B97-D/Def2-TZVP) of dual-hydrogen-bond donor species. Good linear correlation between $\varepsilon_{LUMO}$ and the p$K_a$'s of DHBDs has been found (Fig. S13†).

$V_{buried}$). Despite their ability to enforce a rigid asymmetric environment, which could be beneficial in enantioselective reactions, these catalysts are poorly nucleophilic and predicted to be less reactive. Green and orange species, based on the tetracyclic amino-indanol-derived core developed by Rovis and Bode[127,128] and on morpholine- and pyrrolidine-based triazoliums, have more balanced steric and electronic properties and indeed are among the most popular and versatile NHCs used in organocatalysis.[102] Analysis of the descriptors provided with the 8622 carbenes in OSCAR!(NHC) could eventually be used to tune the catalyst's composition for performance improvement in specific reactions, as outlined in structure–activity–stereoselectivity studies using similar physical organic parameters.[129–131] For example, Rovis, Lee, and co-workers found correlations between the computed gas-phase acidity of a series of triazolium cations and their selectivity in two Umpolung reactions,[132] while Wei and Lan developed a linear model to predict the chemoselectivity of an NHC-catalyzed ester functionalization based on the global nucleophilicity and electrophilicity indices of the species involved in the product-determining step.[133]

In Fig. 6B, each point is coloured according to the nature of the central DHBD unit. Based on $\varepsilon_{LUMO}$, and in agreement with p$K_a$ measurements,[134,135] croconamides and thiosquaramides (purple and red species) are more acidic than thioureas, ureas, and deltamides (yellow, blue, and light blue). Sulfamides (orange points) cover a relatively wider range of $\varepsilon_{LUMO}$ values, implying that the higher electron-withdrawing ability of the sulfonyl group, which should result in stronger acidity of the N–H bonds compared to ureas,[136] is significantly modulated by the substituents. The rapid estimation and comparison of the acidity of various DHBDs is useful for reaction optimization, as dual-hydrogen-bond donors with lower p$K_a$'s have been found to give better enantioselectivities and faster reaction times.[137]

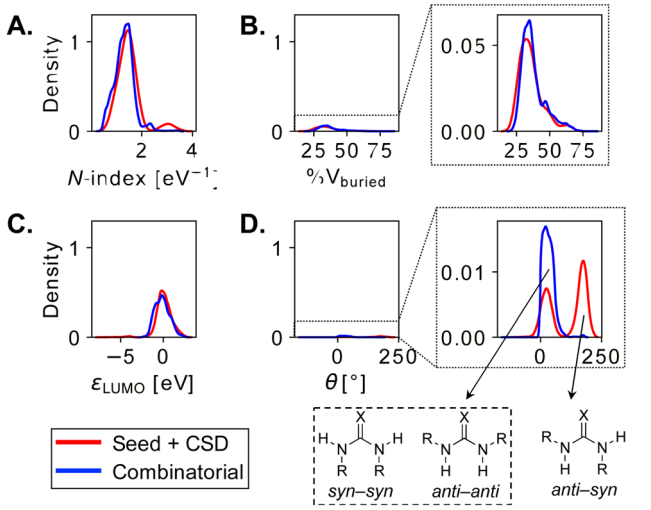


Fig. 7 Distribution plots (*y*-axis: normalized probability density) of molecular descriptors for NHCs (A and B) and DHBDs (C and D) in the seed + CSD-extracted (red curves) and combinatorial databases (blue). X = O/S.

Sulfamides are also the most flexible species, as indicated by the large number of catalysts with $\theta > 80°$. In OSCAR!(DHBD), the majority of structures generated and selected for DFT optimization are in the *anti–anti* or *syn–syn* conformation ($\theta < 80°$, Fig. 7D and S17†),[138] the former being the most relevant to catalysis, since the hydrogens point in the same direction.[139]

If we compare the distribution of $\theta$ values in the "original" and combinatorial datasets (Fig. 7D), we see that many CSD-extracted DHBDs adopt the *anti–syn* conformation ($\theta > 80°$). In a comprehensive study of diaryl(thio)ureas from CSD, Paton *et al.* found that the majority (99%) of ureas exist as *anti–anti* conformers, whereas about 60% thioureas are in the *anti–syn* form.[140] These results agree with our own, with thioureas extracted from CSD having large $\theta$'s (Fig. 7D). The "original" and combinatorial sets are more similarly distributed in terms of the other molecular descriptors (Fig. 7A–C, *N*-index, %$V_{buried}$, and $\varepsilon_{LUMO}$), suggesting that the recombination of the same fragments does not significantly alter the property space covered; instead, the combinatorial strategy provides more instances/structures for each property value.

## Conclusions

We have introduced OSCAR (Organic Structures for CAtalysis Repository), a database of 4000 organocatalysts mined from the literature and CSD and enriched with several thousand species generated from fragments in a combinatorial fashion. We have developed a transferable fragment-based strategy for dataset generation, which exploits the modularity of organocatalysts by defining function-based catalytic motifs and structural substituents. OSCAR covers a wide region of catalyst space with incomparable chemical diversity, and includes a selection of steric and electronic molecular descriptors useful for catalytic properties estimation and performance prediction. All content (geometries, stereoelectronic parameters) is publicly available on the Materials Cloud for interactive visualization with Chemiscope[55] (**https://doi.org/10.24435/materialscloud:gy-3h**) and fully searchable and interoperable with chemoinformatics software (*e.g.*, RDKit, SMILES-based tools); the corresponding chemical space maps could be used for many potential applications, including data and training set curation, organocatalyst inverse design through evolutionary experiments,[56] and mechanistic understanding. We expect OSCAR, and its future extensions and refinements, to assist in the establishment of data-driven and fragment-based reaction optimization methods in organic synthesis.[53]

## Computational methods

### Quantum chemistry

All DFT computations were performed with the Gaussian16 software package.[141] Geometry optimizations were carried out at the B97-D/Def2-TZVP level[142–144] in the gas-phase applying density fitting techniques. ωB97X-D/Def2-TZVP single-point energies[145] were computed in the gas-phase at the B97-D geometries. The ionization potential and electron affinity of a subset 2060 organocatalysts from the seed and CSD datasets

were also computed at the IP/EA-EOM-DLPNO-CCSD[146]/cc-pVTZ level as implemented in Orca 5.0.[147] All coupled cluster computations used the RIJCOSX approximation[148] with the cc-pVTZ/C and the Def2/J auxiliary basis sets for correlation and resolution of identity. This high-level data is available and can be used for the training of ML models. The structures in the combinatorial databases were pre-optimized with the semiempirical GFN2-xTB Hamiltonian[149] in the gas-phase, followed by DFT optimizations and single-points, as described above.

The initial set of Cartesian coordinates for each organocatalyst was either obtained by converting SMILES formats[150] into three-dimensional structures with the 3D structure generator operation (*i.e.*, gen3d operation) implemented in the OpenBabel software,[151] or applying cell2mol[84] on selected CSD entries exported with ConQuest (version 5.42), included in the CCSD software, from the CSD database updated to May 2021. The t-SNE map[91] for the 4000 catalysts in OSCAR was computed on the basis of the FCHL19 representation[92] of each molecule. The perplexity used to generate the structure map was set to 20 and the maximum number of optimization iterations was fixed at 5000.

Open shell single-point computations ($n - 1$ and $n + 1$ electrons) were also performed at the optimized $n$-electron B97-D geometries and uωB97X-D/Def2-TZVP level for the 4000 catalysts in the seed + CSD dataset and for the 8622 carbenes in OSCAR!(NHC). These energies provide an alternative way of estimating the organocatalysts' ionization potential [$\mathrm{IP} = E(n-1) - E(n)$] and electron affinity [$\mathrm{EA} = E(n) - E(n+1)$] (see the ESI† for further details).[152]

### Reaction indices

The organocatalysts' ionization potential (IP) and electron affinity (EA) were estimated from the frontier molecular orbital energies (FMOs) of the $n$-electron species (in the gas-phase, at the ωB97X-D/Def2-TZVP level) using Koopman's theorem[153] within a Hartree–Fock scheme and used to calculate the conceptual DFT descriptors[98,154,155] chemical potential ($\mu$), hardness ($\eta$), $E$-index, $N$-index, and relative nucleophilicity ($N_{\mathrm{rel}}$) as follows:

$$\mu = \frac{(\varepsilon_{\mathrm{LUMO}} + \varepsilon_{\mathrm{HOMO}})}{2} \quad (1)$$

$$\eta = \frac{(\varepsilon_{\mathrm{LUMO}} - \varepsilon_{\mathrm{HOMO}})}{2} \quad (2)$$

$$E\text{-index} = \frac{\mu^2}{2\eta} \quad (3)$$

$$N\text{-index} = \frac{1}{E\text{-index}} \quad (4)$$

$$N_{\mathrm{rel}} = \varepsilon_{\mathrm{HOMO}} - \varepsilon_{\mathrm{HOMO(TCNE)}} \quad (5)$$

where TCNE is tetracyanoethylene. Note that, based on the different formalisms for defining nucleophilicity,[156] a distinction has been made between $N$-index (the reciprocal of the $E$-index) and relative nucleophilicity ($N_{\mathrm{rel}}$).[157]

## Data availability

The structures of the the organocatalysts and their stereoelectronic descriptors are publicly available on the Materials Cloud for interactive visualization with Chemiscope (**https://archive.materialscloud.org/record/2022.106**).

## Author contributions

S. G. and C. C. conceived the project. S. G. performed DFT computations, curated the data, and analysed the results, with help from P. v. G. R. L. developed and implemented scripts for the combinatorial databases generation and analysis. S. V. developed and implemented cell2mol. A. F. provided support with the coupled cluster computations and t-SNE plot generation. S. G. and C. C. wrote the manuscript with help and feedback from all authors. C. C. provided supervision throughout.

## Conflicts of interest

There are no conflicts to declare.

## Acknowledgements

S. G. acknowledges the European Research Council (ERC, Grant Agreement No. 817977) within the framework of European Union's H2020 for financial support. The National Centre of Competence in Research (NCCR) "Sustainable chemical process through catalysis (Catalysis)" of the Swiss National Science Foundation (SNSF, grant number 180544) is acknowledged for financial support of P. v. G. and R. L. S. V. and A. F. acknowledge the National Centre of Competence in Research (NCCR) "Materials' Revolution: Computational Design and Discovery of Novel Materials (MARVEL)" of the Swiss National Science Foundation (SNSF, grant number 182892). The authors also acknowledge support from EPFL. Dr Guillaume Fraux is acknowledged for his help with Chemiscope.

## References

1 C. Bo, F. Maseras and N. López, The role of computational results databases in accelerating the discovery of catalysts, *Nat. Catal.*, 2018, **1**, 809–810.
2 A. Nandy, C. Duan and H. J. Kulik, Audacity of huge: overcoming challenges of data scarcity and data quality for machine learning in computational materials discovery, *Curr. Opin. Chem. Eng.*, 2022, **36**, 100778.
3 A. McNally, C. K. Prier and D. W. C. MacMillan, Discovery of an α-Amino C–H Arylation Reaction Using the Strategy of Accelerated Serendipity, *Science*, 2011, **334**, 1114–1117.
4 S. Mitsumori, H. Zhang, P. Ha-Yeon Cheong, K. N. Houk, F. Tanaka and C. F. Barbas, Direct Asymmetric anti-Mannich-Type Reactions Catalyzed by a Designed Amino Acid, *J. Am. Chem. Soc.*, 2006, **128**, 1040–1041.

5 E. M. Fleming, C. Quigley, I. Rozas and S. J. Connon, Computational Study-Led Organocatalyst Design: A Novel, Highly Active Urea-Based Catalyst for Addition Reactions to Epoxides, *J. Org. Chem.*, 2008, **73**, 948–956.
6 A. D. Gammack Yamagata, S. Datta, K. E. Jackson, L. Stegbauer, R. S. Paton and D. J. Dixon, Enantioselective Desymmetrization of Prochiral Cyclohexanones by Organocatalytic Intramolecular Michael Additions to α,β-Unsaturated Esters, *Angew. Chem., Int. Ed.*, 2015, **54**, 4899–4903.
7 I. Iribarren and C. Trujillo, Improving phase-transfer catalysis by enhancing non-covalent interactions, *Phys. Chem. Chem. Phys.*, 2020, **22**, 21015–21021.
8 A. R. Rosales, J. Wahlers, E. Limé, R. E. Meadows, K. W. Leslie, R. Savin, F. Bell, E. Hansen, P. Helquist, R. H. Munday, O. Wiest and P.-O. Norrby, Rapid virtual screening of enantioselective catalysts using CatVS, *Nat. Catal.*, 2019, **2**, 41–45.
9 A. C. Doney, B. J. Rooks, T. Lu and S. E. Wheeler, Design of Organocatalysts for Asymmetric Propargylations through Computational Screening, *ACS Catal.*, 2016, **6**, 7948–7955.
10 G. G. Gerosa, R. A. Spanevello, A. G. Suárez and A. M. Sarotti, Joint Experimental, *in Silico*, and NMR Studies toward the Rational Design of Iminium-Based Organocatalyst Derived from Renewable Sources, *J. Org. Chem.*, 2015, **80**, 7626–7634.
11 G. G. Gerosa, M. O. Marcarino, R. A. Spanevello, A. G. Suárez and A. M. Sarotti, Re-Engineering Organocatalysts for Asymmetric Friedel–Crafts Alkylation of Indoles through Computational Studies, *J. Org. Chem.*, 2020, **85**, 9969–9978.
12 M. Foscato and V. R. Jensen, Automated *in Silico* Design of Homogeneous Catalysts, *ACS Catal.*, 2020, **10**, 2354–2377.
13 K. N. Houk and F. Liu, Holy Grails for Computational Organic Chemistry and Biochemistry, *Acc. Chem. Res.*, 2017, **50**, 539–543.
14 L. Falivene, Z. Cao, A. Petta, L. Serra, A. Poater, R. Oliva, V. Scarano and L. Cavallo, Towards the online computer-aided design of catalytic pockets, *Nat. Chem.*, 2019, **11**, 872–879.
15 A. Nandy, C. Duan, M. G. Taylor, F. Liu, A. H. Steeves and H. J. Kulik, Computational Discovery of Transition-metal Complexes: From High-throughput Screening to Machine Learning, *Chem. Rev.*, 2021, **121**, 9927–10000.
16 J. P. Janet, C. Duan, A. Nandy, F. Liu and H. J. Kulik, Navigating Transition-Metal Chemical Space: Artificial Intelligence for First-Principles Design, *Acc. Chem. Res.*, 2021, **54**, 532–545.
17 A. Nandy, J. Zhu, J. P. Janet, C. Duan, R. B. Getman and H. J. Kulik, Machine Learning Accelerates the Discovery of Design Rules and Exceptions in Stable Metal–Oxo Intermediate Formation, *ACS Catal.*, 2019, **9**, 8243–8255.
18 S. Gugler, J. P. Janet and H. J. Kulik, Enumeration of *de novo* inorganic complexes for chemical discovery and machine learning, *Mol. Syst. Des. Eng.*, 2020, **5**, 139–152.
19 F. Liu, C. Duan and H. J. Kulik, Rapid Detection of Strong Correlation with Machine Learning for Transition-Metal Complex High-Throughput Screening, *J. Phys. Chem. Lett.*, 2020, **11**, 8067–8076.
20 J. A. Hueffel, T. Sperger, I. Funes-Ardoiz, J. S. Ward, K. Rissanen and F. Schoenebeck, Accelerated dinuclear palladium catalyst identification through unsupervised machine learning, *Science*, 2021, **374**, 1134–1140.
21 N. Fey, A. C. Tsipis, S. E. Harris, J. N. Harvey, A. G. Orpen and R. A. Mansson, Development of a Ligand Knowledge Base, Part 1: Computational Descriptors for Phosphorus Donor Ligands, *Chem.–Eur. J.*, 2006, **12**, 291–302.
22 J. Jover, N. Fey, J. N. Harvey, G. C. Lloyd-Jones, A. G. Orpen, G. J. J. Owen-Smith, P. Murray, D. R. J. Hose, R. Osborne and M. Purdie, Expansion of the Ligand Knowledge Base for Monodentate P-Donor Ligands (LKB-P), *Organometallics*, 2010, **29**, 6245–6258.
23 N. Fey, J. N. Harvey, G. C. Lloyd-Jones, P. Murray, A. G. Orpen, R. Osborne and M. Purdie, Computational Descriptors for Chelating P,P- and P,N-Donor Ligands1, *Organometallics*, 2008, **27**, 1372–1383.
24 D. J. Durand and N. Fey, Computational Ligand Descriptors for Catalyst Design, *Chem. Rev.*, 2019, **119**, 6561–6594.
25 D. J. Durand and N. Fey, Building a Toolbox for the Analysis and Prediction of Ligand and Catalyst Effects in Organometallic Catalysis, *Acc. Chem. Res.*, 2021, **54**, 837–848.
26 N. Fey, A. Koumi, A. V. Malkov, J. D. Moseley, B. N. Nguyen, S. N. G. Tyler and C. E. Willans, Mapping the properties of bidentate ligands with calculated descriptors (LKB-bid), *Dalton Trans.*, 2020, **49**, 8169–8178.
27 T. Gensch, G. dos Passos Gomes, P. Friederich, E. Peters, T. Gaudin, R. Pollice, K. Jorner, A. Nigam, M. Lindner-D'Addario, M. S. Sigman and A. Aspuru-Guzik, A Comprehensive Discovery Platform for Organophosphorus Ligands for Catalysis, *J. Am. Chem. Soc.*, 2022, **144**, 1205–1217.
28 M. Foscato, V. Venkatraman, G. Occhipinti, B. K. Alsberg and V. R. Jensen, Automated Building of Organometallic Complexes from 3D Fragments, *J. Chem. Inf. Model.*, 2014, **54**, 1919–1931.
29 M. Foscato, G. Occhipinti, V. Venkatraman, B. K. Alsberg and V. R. Jensen, Automated Design of Realistic Organometallic Molecules from Fragments, *J. Chem. Inf. Model.*, 2014, **54**, 767–780.
30 Y. Chu, W. Heyndrickx, G. Occhipinti, V. R. Jensen and B. K. Alsberg, An Evolutionary Algorithm for *de Novo* Optimization of Functional Transition Metal Compounds, *J. Am. Chem. Soc.*, 2012, **134**, 8885–8895.
31 T. W. Thorpe, J. R. Marshall, V. Harawa, R. E. Ruscoe, A. Cuetos, J. D. Finnigan, A. Angelastro, R. S. Heath, F. Parmeggiani, S. J. Charnock, R. M. Howard, R. Kumar, D. S. B. Daniels, G. Grogan and N. J. Turner, Multifunctional biocatalyst for conjugate reduction and reductive amination, *Nature*, 2022, **604**, 86–91.
32 G. Lapidoth, O. Khersonsky, R. Lipsh, O. Dym, S. Albeck, S. Rogotner and S. J. Fleishman, Highly active enzymes by automated combinatorial backbone assembly and sequence design, *Nat. Commun.*, 2018, **9**, 2780.

33 T. P. Yoon and E. N. Jacobsen, Privileged Chiral Catalysts, *Science*, 2003, **299**, 1691–1693.
34 D. A. Strassfeld, R. F. Algera, Z. K. Wickens and E. N. Jacobsen, A Case Study in Catalyst Generality: Simultaneous, Highly-Enantioselective Brønsted- and Lewis-Acid Mechanisms in Hydrogen-Bond-Donor Catalyzed Oxetane Openings, *J. Am. Chem. Soc.*, 2021, **143**, 9585–9594.
35 J. Seayad and B. List, Asymmetric Organocatalysis, *Org. Biomol. Chem.*, 2005, **3**, 719–724.
36 R. Kenny and F. Liu, Trifunctional Organocatalysts: Catalytic Proficiency by Cooperative Activation, *Eur. J. Org. Chem.*, 2015, **2015**, 5304–5319.
37 P. Kirkpatrick and C. Ellis, Chemical space, *Nature*, 2004, **432**, 823.
38 J.-L. Reymond, The Chemical Space Project, *Acc. Chem. Res.*, 2015, **48**, 722–730.
39 G. Schneider and U. Fechner, Computer-based *de novo* design of drug-like molecules, *Nat. Rev. Drug Discovery*, 2005, **4**, 649–663.
40 T. Liu, M. Naderi, C. Alvin, S. Mukhopadhyay and M. Brylinski, Break Down in Order To Build Up: Decomposing Small Molecules for Fragment-Based Drug Design with *e*MolFrag, *J. Chem. Inf. Model.*, 2017, **57**, 627–631.
41 I. O. Betinol, Y. Kuang and J. P. Reid, Guiding Target Synthesis with Statistical Modeling Tools: A Case Study in Organocatalysis, *Org. Lett.*, 2022, **24**, 1429–1433.
42 J. P. Reid, K. Ermanis and J. M. Goodman, BINOPtimal: a web tool for optimal chiral phosphoric acid catalyst selection, *Chem. Commun.*, 2019, **55**, 1778–1781.
43 Y. Guan, V. M. Ingman, B. J. Rooks and S. E. Wheeler, AARON: An Automated Reaction Optimizer for New Catalysts, *J. Chem. Theory Comput.*, 2018, **14**, 5249–5261.
44 N. Weill, C. R. Corbeil, J. W. De Schutter and N. Moitessier, Toward a computational tool predicting the stereochemical outcome of asymmetric reactions: Development of the molecular mechanics-based program ACE and application to asymmetric epoxidation reactions, *J. Comput. Chem.*, 2011, **32**, 2878–2889.
45 M. Burai Patrascu, J. Pottel, S. Pinus, M. Bezanson, P.-O. Norrby and N. Moitessier, From desktop to benchtop with automated computational workflows for computer-aided design in asymmetric catalysis, *Nat. Catal.*, 2020, **3**, 574–584.
46 A. F. Zahrt, J. J. Henle, B. T. Rose, Y. Wang, W. T. Darrow and S. E. Denmark, Prediction of higher-selectivity catalysts by computer-driven workflow and machine learning, *Science*, 2019, **363**, eaau5631.
47 J. P. Reid and M. S. Sigman, Holistic prediction of enantioselectivity in asymmetric catalysis, *Nature*, 2019, **571**, 343–348.
48 J. P. Reid, L. Simón and J. M. Goodman, A Practical Guide for Predicting the Stereochemistry of Bifunctional Phosphoric Acid Catalyzed Reactions of Imines, *Acc. Chem. Res.*, 2016, **49**, 1029–1041.
49 A. Shoja, J. Zhai and J. P. Reid, Comprehensive Stereochemical Models for Selectivity Prediction in Diverse Chiral Phosphate-Catalyzed Reaction Space, *ACS Catal.*, 2021, **11**, 11897–11905.
50 J. Werth and M. S. Sigman, Connecting and Analyzing Enantioselective Bifunctional Hydrogen Bond Donor Catalysis Using Data Science Tools, *J. Am. Chem. Soc.*, 2020, **142**, 16382–16391.
51 K. W. Lexa, K. M. Belyk, J. Henle, B. Xiang, R. P. Sheridan, S. E. Denmark, R. T. Ruck and E. C. Sherer, Application of Machine Learning and Reaction Optimization for the Iterative Improvement of Enantioselectivity of Cinchona-Derived Phase Transfer Catalysts, *Org. Process Res. Dev.*, 2022, **26**, 670–682.
52 S. Gallarati, R. Fabregat, R. Laplaza, S. Bhattacharjee, M. D. Wodrich and C. Corminboeuf, Reaction-based machine learning representations for predicting the enantioselectivity of organocatalysts, *Chem. Sci.*, 2021, **12**, 6879–6889.
53 S. Gallarati, R. Laplaza and C. Corminboeuf, Harvesting the fragment-based nature of bifunctional organocatalysts to enhance their activity, *Org. Chem. Front.*, 2022, **9**, 4041–4051.
54 N. Tsuji, P. Sidorov, C. Zhu, Y. Nagata, T. Gimadiev, A. Varnek and B. List, Predicting Highly Enantioselective Catalysts Using Tunable Fragment Descriptors, *ChemRxiv*, 2022, DOI: **10.26434/chemrxiv-2022-bsmdl**.
55 G. Fraux, R. K. Cersonsky and M. Ceriotti, Chemiscope: interactive structure–property explorer for materials and molecules, *J. Open Source Softw.*, 2020, **5**, 2117.
56 R. Laplaza, S. Gallarati and C. Corminboeuf, Genetic Optimization of Homogeneous Catalysts, *Chem.: Methods*, 2022, e202100107.
57 M. C. Holland and R. Gilmour, Deconstructing Covalent Organocatalysis, *Angew. Chem., Int. Ed.*, 2015, **54**, 3862–3871.
58 A. Dondoni and A. Massi, Asymmetric Organocatalysis: From Infancy to Adolescence, *Angew. Chem., Int. Ed.*, 2008, **47**, 4638–4660.
59 P. I. Dalko and L. Moisan, In the Golden Age of Organocatalysis, *Angew. Chem., Int. Ed.*, 2004, **43**, 5138–5175.
60 P. I. Dalko and L. Moisan, Enantioselective Organocatalysis, *Angew. Chem., Int. Ed.*, 2001, **40**, 3726–3748.
61 R. Maji, S. C. Mallojjala and S. E. Wheeler, Chiral phosphoric acid catalysis: from numbers to insights, *Chem. Soc. Rev.*, 2018, **47**, 1142–1158.
62 D. Enders, O. Niemeier and A. Henseler, Organocatalysis by N-Heterocyclic Carbenes, *Chem. Rev.*, 2007, **107**, 5606–5655.
63 O. A. Wong and Y. Shi, Organocatalytic Oxidation. Asymmetric Epoxidation of Olefins Catalyzed by Chiral Ketones and Iminium Salts, *Chem. Rev.*, 2008, **108**, 3958–3987.
64 E. M. McGarrigle, E. L. Myers, O. Illa, M. A. Shaw, S. L. Riches and V. K. Aggarwal, Chalcogenides as Organocatalysts, *Chem. Rev.*, 2007, **107**, 5841–5883.

65 T. Marcelli and H. Hiemstra, Cinchona Alkaloids in Asymmetric Organocatalysis, *Synthesis*, 2010, 1229–1279.
66 M. Benaglia and S. Rossi, Chiral phosphine oxides in present-day organocatalysis, *Org. Biomol. Chem.*, 2010, **8**, 3824–3830.
67 X. Liu, L. Lin and X. Feng, Chiral *N*,*N*′-Dioxides: New Ligands and Organocatalysts for Catalytic Asymmetric Reactions, *Acc. Chem. Res.*, 2011, **44**, 574–587.
68 Y. Wei and M. Shi, Applications of Chiral Phosphine-Based Organocatalysts in Catalytic Asymmetric Reactions, *Chem.–Asian J.*, 2014, **9**, 2720–2734.
69 Q. Yang, Y. Li, J.-D. Yang, Y. Liu, L. Zhang, S. Luo and J.-P. Cheng, Holistic Prediction of the p$K_a$ in Diverse Solvents Based on a Machine-Learning Approach, *Angew. Chem., Int. Ed.*, 2020, **59**, 19282–19291.
70 C. Yang, X.-S. Xue, X. Li and J.-P. Cheng, Computational Study on the Acidic Constants of Chiral Brønsted Acids in Dimethyl Sulfoxide, *J. Org. Chem.*, 2014, **79**, 4340–4351.
71 P. Christ, A. G. Lindsay, S. S. Vormittag, J.-M. Neudörfl, A. Berkessel and A. C. O'Donoghue, p$K_a$ Values of Chiral Brønsted Acid Catalysts: Phosphoric Acids/Amides, Sulfonyl/Sulfuryl Imides, and Perfluorinated TADDOLs (TEFDDOLs), *Chem.–Eur. J.*, 2011, **17**, 8524–8528.
72 R. R. Walvoord, P. N. H. Huynh and M. C. Kozlowski, Quantification of Electrophilic Activation by Hydrogen-Bonding Organocatalysts, *J. Am. Chem. Soc.*, 2014, **136**, 16055–16065.
73 A. Moyano, in *Stereoselective Organocatalysis*, John Wiley & Sons, Ltd, 2013, pp. 11–80.
74 P. I. Dalko, in *Enantioselective Organocatalysis*, John Wiley & Sons, Ltd, 2007, pp. 1–17.
75 Y. R. Chi, Comprehensive Enantioselective Organocatalysis. Edited by Peter I. Dalko, *Angew. Chem., Int. Ed.*, 2014, **53**, 6858.
76 B. List, *Asymmetric Organocatalysis*, Springer, 2009.
77 P. Pihko, in *Hydrogen Bonding in Organic Synthesis*, John Wiley & Sons, Ltd, 2009, pp. 1–4.
78 *Asymmetric Organocatalysis*, ChemFiles, Sigma-Aldrich, 2006, vol. 6, pp. 1–16.
79 Organocatalysis, ChemFiles, Sigma-Aldrich, 2007, vol. 7, pp. 1–24.
80 S.-H. Xiang and B. Tan, Advances in asymmetric organocatalysis over the last 10 years, *Nat. Commun.*, 2020, **11**, 3786.
81 C. R. Groom and F. H. Allen, The Cambridge Structural Database in Retrospect and Prospect, *Angew. Chem., Int. Ed.*, 2014, **53**, 662–671.
82 C. R. Groom, I. J. Bruno, M. P. Lightfoot and S. C. Ward, The Cambridge Structural Database, *Acta Crystallogr., Sect. B: Struct. Sci., Cryst. Eng. Mater.*, 2016, **72**, 171–179.
83 S. Gražulis, D. Chateigner, R. T. Downs, A. F. T. Yokochi, M. Quirós, L. Lutterotti, E. Manakova, J. Butkus, P. Moeck and A. Le Bail, Crystallography Open Database – an open-access collection of crystal structures, *J. Appl. Crystallogr.*, 2009, **42**, 726–729.
84 S. Vela, R. Laplaza, Y. Cho and C. Corminboeuf, Cell2mol: encoding chemistry to interpret crystallographic data, *npj Comput. Mater.*, 2022, **8**, 188.
85 L. Garuti, M. Roberti, G. Bottegoni and M. Ferraro, Diaryl Urea: A Privileged Structure in Anticancer Agents, *Curr. Med. Chem.*, 2016, **23**, 1528–1548.
86 S. M. Anil, N. Rajeev, K. R. Kiran, T. R. Swaroop, N. Mallesha, R. Shobith and M. P. Sadashiva, Multi-pharmacophore Approach to Bio-therapeutics: Piperazine Bridged Pseudo-peptidic Urea/Thiourea Derivatives as Anti-oxidant Agents, *Int. J. Pept. Res. Ther.*, 2020, **26**, 151–158.
87 S. Azeem, A. Ataf Ali, Q. Ashfaq Mahmood and B. Amin, Thiourea Derivatives in Drug Design and Medicinal Chemistry: A Short Review, *J. Drug Des. Med. Chem.*, 2016, **2**, 10–20.
88 V. B. Bregović, N. Basarić and K. Mlinarić-Majerski, Anion binding with urea and thiourea derivatives, *Coord. Chem. Rev.*, 2015, **295**, 80–124.
89 L.-W. Xu, J. Luo and Y. Lu, Asymmetric catalysis with chiral primary amine-based organocatalysts, *Chem. Commun.*, 2009, 1807–1821.
90 G. Li Petri, M. V. Raimondi, V. Spanò, R. Holl, P. Barraja and A. Montalbano, Pyrrolidine in Drug Discovery: A Versatile Scaffold for Novel Biologically Active Compounds, *Top. Curr. Chem.*, 2021, **379**, 34.
91 L. van der Maaten and G. Hinton, Visualizing Data using t-SNE, *J. Mach. Learn. Res.*, 2008, **9**, 2579–2605.
92 A. S. Christensen, L. A. Bratholm, F. A. Faber and O. Anatole von Lilienfeld, FCHL revisited: Faster and more accurate quantum machine learning, *J. Chem. Phys.*, 2020, **152**, 044107.
93 T. N. Nguyen, P.-A. Chen, K. Setthakarn and J. A. May, Chiral Diol-Based Organocatalysts in Enantioselective Reactions, *Molecules*, 2018, **23**, 2317.
94 D. Parmar, E. Sugiono, S. Raja and M. Rueping, Complete Field Guide to Asymmetric BINOL-Phosphate Derived Brønsted Acid and Metal Catalysis: History and Classification by Mode of Activation; Brønsted Acidity, Hydrogen Bonding, Ion Pairing, and Metal Phosphates, *Chem. Rev.*, 2014, **114**, 9047–9153.
95 L. Shu and Y. Shi, An Efficient Ketone-Catalyzed Epoxidation Using Hydrogen Peroxide as Oxidant, *J. Org. Chem.*, 2000, **65**, 8807–8810.
96 S. E. Denmark and Z. Wu, The Development of Chiral, Nonracemic Dioxiranes for the Catalytic, Enantioselective Epoxidation of Alkenes, *Synlett*, 2000, **1999**, 847–859.
97 M. Formica, D. Rozsar, G. Su, A. J. M. Farley and D. J. Dixon, Bifunctional Iminophosphorane Superbase Catalysis: Applications in Organic Synthesis, *Acc. Chem. Res.*, 2020, **53**, 2235–2247.
98 D. Chakraborty and P. K. Chattaraj, Conceptual density functional theory based electronic structure principles, *Chem. Sci.*, 2021, **12**, 6264–6279.
99 B. Lee, J. Yoo and K. Kang, Predicting the chemical reactivity of organic materials using a machine-learning approach, *Chem. Sci.*, 2020, **11**, 7813–7822.

100 F. W. Patureau, C. Worch, M. A. Siegler, A. L. Spek, C. Bolm and J. N. H. Reek, SIAPhos: Phosphorylated Sulfonimidamides and their Use in Iridium-Catalyzed Asymmetric Hydrogenations of Sterically Hindered Cyclic Enamides, *Adv. Synth. Catal.*, 2012, **354**, 59–64.

101 B. Xiang, K. M. Belyk, R. A. Reamer and N. Yasuda, Discovery and Application of Doubly Quaternized Cinchona-Alkaloid-Based Phase-Transfer Catalysts, *Angew. Chem., Int. Ed.*, 2014, **53**, 8375–8378.

102 D. M. Flanigan, F. Romanov-Michailidis, N. A. White and T. Rovis, Organocatalytic Reactions Enabled by N-Heterocyclic Carbenes, *Chem. Rev.*, 2015, **115**, 9307–9387.

103 A. Mood, M. Tavakoli, E. Gutman, D. Kadish, P. Baldi and D. L. Van Vranken, Methyl Anion Affinities of the Canonical Organic Functional Groups, *J. Org. Chem.*, 2020, **85**, 4096–4102.

104 D. Kadish, A. D. Mood, M. Tavakoli, E. S. Gutman, P. Baldi and D. L. Van Vranken, Methyl Cation Affinities of Canonical Organic Functional Groups, *J. Org. Chem.*, 2021, **86**, 3721–3729.

105 K. Kaupmees, N. Tolstoluzhsky, S. Raja, M. Rueping and I. Leito, On the Acidity and Reactivity of Highly Effective Chiral Brønsted Acid Catalysts: Establishment of an Acidity Scale, *Angew. Chem., Int. Ed.*, 2013, **52**, 11569–11572.

106 G. Jakab, C. Tancon, Z. Zhang, K. M. Lippert and P. R. Schreiner, (Thio)urea Organocatalyst Equilibrium Acidities in DMSO, *Org. Lett.*, 2012, **14**, 1724–1727.

107 X. Ni, X. Li, Z. Wang and J.-P. Cheng, Squaramide Equilibrium Acidities in DMSO, *Org. Lett.*, 2014, **16**, 1786–1789.

108 Z. Li, X. Li, X. Ni and J.-P. Cheng, Equilibrium Acidities of Proline Derived Organocatalysts in DMSO, *Org. Lett.*, 2015, **17**, 1196–1199.

109 Y. Li, L. Zhang and S. Luo, Bond Energies of Enamines, *ACS Omega*, 2022, **7**, 6354–6374.

110 B. Maji, M. Breugst and H. Mayr, N-Heterocyclic Carbenes: Organocatalysts with Moderate Nucleophilicity but Extraordinarily High Lewis Basicity, *Angew. Chem., Int. Ed.*, 2011, **50**, 6915–6919.

111 H. Mayr, S. Lakhdar, B. Maji and A. R. Ofial, A quantitative approach to nucleophilic organocatalysis, *Beilstein J. Org. Chem.*, 2012, **8**, 1458–1478.

112 F. An, B. Maji, E. Min, A. R. Ofial and H. Mayr, Basicities and Nucleophilicities of Pyrrolidines and Imidazolidinones Used as Organocatalysts, *J. Am. Chem. Soc.*, 2020, **142**, 1526–1547.

113 B. Maji, C. Joannesse, T. A. Nigst, A. D. Smith and H. Mayr, Nucleophilicities and Lewis Basicities of Isothiourea Derivatives, *J. Org. Chem.*, 2011, **76**, 5104–5112.

114 T. Sander, J. Freyss, M. von Korff and C. Rufener, DataWarrior: An Open-Source Program For Chemistry Aware Data Visualization And Analysis, *J. Chem. Inf. Model.*, 2015, **55**, 460–473.

115 D. Enders and T. Balensiefer, Nucleophilic Carbenes in Asymmetric Organocatalysis, *Acc. Chem. Res.*, 2004, **37**, 534–541.

116 M. N. Hopkinson, C. Richter, M. Schedler and F. Glorius, An overview of N-heterocyclic carbenes, *Nature*, 2014, **510**, 485–496.

117 A. T. Biju, N. Kuhl and F. Glorius, Extending NHC-Catalysis: Coupling Aldehydes with Unconventional Reaction Partners, *Acc. Chem. Res.*, 2011, **44**, 1182–1195.

118 S. J. Ryan, L. Candish and D. W. Lupton, Acyl anion free N-heterocyclic carbene organocatalysis, *Chem. Soc. Rev.*, 2013, **42**, 4906–4917.

119 RDKit: *Open-Source Chemoinformatics and Machine Learning*. **https://www.rdkit.org**.

120 J. Dotson, L. van Dijk, J. Timmerman, S. Grosslight, R. Walroth, K. Püntener, F. Gosselin, K. Mack and M. S. Sigman, Data-driven multi-objective optimization tactics for catalytic asymmetric reactions, *ChemRxiv*, 2022.

121 H. Clavier and S. P. Nolan, Percent buried volume for phosphine and N-heterocyclic carbene ligands: steric properties in organometallic chemistry, *Chem. Commun.*, 2010, **46**, 841–861.

122 A. Gómez-Suárez, D. J. Nelson and S. P. Nolan, Quantifying and understanding the steric properties of N-heterocyclic carbenes, *Chem. Commun.*, 2017, **53**, 2650–2660.

123 Z. Li, X. Li and J.-P. Cheng, An Acidity Scale of Triazolium-Based NHC Precursors in DMSO, *J. Org. Chem.*, 2017, **82**, 9675–9681.

124 X.-Y. Chen, Z.-H. Gao and S. Ye, Bifunctional N-Heterocyclic Carbenes Derived from l-Pyroglutamic Acid and Their Applications in Enantioselective Organocatalysis, *Acc. Chem. Res.*, 2020, **53**, 690–702.

125 Y.-R. Zhang, L. He, X. Wu, P.-L. Shao and S. Ye, Chiral N-Heterocyclic Carbene Catalyzed Staudinger Reaction of Ketenes with Imines: Highly Enantioselective Synthesis of N-Boc β-Lactams, *Org. Lett.*, 2008, **10**, 277–280.

126 L. He, Y.-R. Zhang, X.-L. Huang and S. Ye, Chiral Bifunctional N-Heterocyclic Carbenes: Synthesis and Application in the Aza-Morita-Baylis-Hillman Reaction, *Synthesis*, 2008, **2008**, 2825–2829.

127 M. S. Kerr, J. Read de Alaniz and T. Rovis, A Highly Enantioselective Catalytic Intramolecular Stetter Reaction, *J. Am. Chem. Soc.*, 2002, **124**, 10298–10299.

128 M. He, J. R. Struble and J. W. Bode, Highly Enantioselective Azadiene Diels–Alder Reactions Catalyzed by Chiral N-Heterocyclic Carbenes, *J. Am. Chem. Soc.*, 2006, **128**, 8418–8420.

129 N. Wang, J. Xu and J. K. Lee, The importance of N-heterocyclic carbene basicity in organocatalysis, *Org. Biomol. Chem.*, 2018, **16**, 8230–8244.

130 Z. Li, X. Li and J.-P. Cheng, Recent Progress in Equilibrium Acidity Studies of Organocatalysts, *Synlett*, 2019, **30**, 1940–1949.

131 S. C. Gadekar, V. Dhayalan, A. Nandi, I. L. Zak, M. S. Mizrachi, S. Kozuch and A. Milo, Rerouting the Organocatalytic Benzoin Reaction toward Aldehyde Deuteration, *ACS Catal.*, 2021, **11**, 14561–14569.

132 Y. Niu, N. Wang, A. Muñoz, J. Xu, H. Zeng, T. Rovis and J. K. Lee, Experimental and Computational Gas Phase Acidities of Conjugate Acids of Triazolylidene Carbenes:

Rationalizing Subtle Electronic Effects, *J. Am. Chem. Soc.*, 2017, **139**, 14917–14930.

133 X. Li, J. Xu, S.-J. Li, L.-B. Qu, Z. Li, Y. R. Chi, D. Wei and Y. Lan, Prediction of NHC-catalyzed chemoselective functionalizations of carbonyl compounds: a general mechanistic map, *Chem. Sci.*, 2020, **11**, 7214–7225.

134 J. Ho, V. E. Zwicker, K. K. Y. Yuen and K. A. Jolliffe, Quantum Chemical Prediction of Equilibrium Acidities of Ureas, Deltamides, Squaramides, and Croconamides, *J. Org. Chem.*, 2017, **82**, 10732–10736.

135 V. E. Zwicker, K. K. Y. Yuen, D. G. Smith, J. Ho, L. Qin, P. Turner and K. A. Jolliffe, Deltamides and Croconamides: Expanding the Range of Dual H-bond Donors for Selective Anion Recognition, *Chem.–Eur. J.*, 2018, **24**, 1140–1150.

136 X. Zhang, S. Liu, X. Li, M. Yan and A. S. C. Chan, Highly enantioselective conjugate addition of aldehydes to nitroolefins catalyzed by chiral bifunctional sulfamides, *Chem. Commun.*, 2009, 833–835.

137 X. Li, H. Deng, B. Zhang, J. Li, L. Zhang, S. Luo and J.-P. Cheng, Physical Organic Study of Structure-Activity-Enantioselectivity Relationships in Asymmetric Bifunctional Thiourea Catalysis: Hints for the Design of New Organocatalysts, *Chem.–Eur. J.*, 2010, **16**, 450–455.

138 I. Sandler, F. A. Larik, N. Mallo, J. E. Beves and J. Ho, Anion Binding Affinity: Acidity *versus* Conformational Effects, *J. Org. Chem.*, 2020, **85**, 8074–8084.

139 A. Wittkopp and P. R. Schreiner, Metal-Free, Noncovalent Catalysis of Diels–Alder Reactions by Neutral Hydrogen Bond Donors in Organic Solvents and in Water, *Chem.–Eur. J.*, 2003, **9**, 407–414.

140 G. Luchini, D. M. H. Ascough, J. V. Alegre-Requena, V. Gouverneur and R. S. Paton, Data-mining the diaryl(thio)urea conformational landscape: understanding the contrasting behavior of ureas and thioureas with quantum chemistry, *Tetrahedron*, 2019, **75**, 697–702.

141 M. Frisch, G. Trucks, H. Schlegel, G. Scuseria, M. Robb, J. Cheeseman, J. Montgomery, T. Vreven, K. Kudin, J. Burant, J. Millam, S. Iyengar, J. Tomasi, V. Barone, B. Mennucci, M. Cossi, G. Scalmani, N. Rega, G. Petersson, H. Nakatsuji, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, M. Klene, X. Li, J. Knox, H. Hratchian, J. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. Stratmann, O. Yazyev, A. Austin, R. Cammi, C. Pomelli, J. Ochterski, P. Ayala, K. Morokuma, G. Voth, P. Salvador, J. Dannenberg, V. Zakrzewski, S. Dapprich, A. Daniels, M. Strain, O. Farkas, D. Malick, A. Rabuck, K. Raghavachari, J. Foresman, J. Ortiz, Q. Cui, A. Baboul, S. Clifford, J. Cioslowski, B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R. Martin, D. Fox, T. Keith, A. Laham, C. Peng, A. Nanayakkara, M. Challacombe, P. Gill, B. Johnson, W. Chen, M. Wong, C. Gonzalez and J. Pople, *Gaussian 16*, Revision C.01, Wallingford, CT, 2016.

142 A. D. Becke, Density-functional thermochemistry. V. Systematic optimization of exchange–correlation functionals, *J. Chem. Phys.*, 1997, **107**, 8554–8560.

143 S. Grimme, Semiempirical GGA-type density functional constructed with a long-range dispersion correction, *J. Comput. Chem.*, 2006, **27**, 1787–1799.

144 F. Weigend and R. Ahlrichs, Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy, *Phys. Chem. Chem. Phys.*, 2005, **7**, 3297–3305.

145 J.-D. Chai and M. Head-Gordon, Systematic optimization of long-range corrected hybrid density functionals, *J. Chem. Phys.*, 2008, **128**, 084106.

146 S. Haldar, C. Riplinger, B. Demoulin, F. Neese, R. Izsak and A. K. Dutta, Multilayer Approach to the IP-EOM-DLPNO-CCSD Method: Theory, Implementation, and Application, *J. Chem. Theory Comput.*, 2019, **15**, 2265–2277.

147 F. Neese, F. Wennmohs, U. Becker and C. Riplinger, The ORCA quantum chemistry program package, *J. Chem. Phys.*, 2020, **152**, 224108.

148 F. Neese, F. Wennmohs, A. Hansen and U. Becker, Efficient, approximate and parallel Hartree–Fock and hybrid DFT calculations. A 'chain-of-spheres' algorithm for the Hartree–Fock exchange, *Chem. Phys.*, 2009, **356**, 98–109.

149 C. Bannwarth, S. Ehlert and S. Grimme, GFN2-xTB—An Accurate and Broadly Parametrized Self-Consistent Tight-Binding Quantum Chemical Method with Multipole Electrostatics and Density-Dependent Dispersion Contributions, *J. Chem. Theory Comput.*, 2019, **15**, 1652–1671.

150 D. Weininger, A. Weininger and J. L. Weininger, SMILES. 2. Algorithm for generation of unique SMILES notation, *J. Chem. Inf. Comput. Sci.*, 1989, **29**, 97–101.

151 N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch and G. R. Hutchison, Open Babel: An open chemical toolbox, *J. Cheminf.*, 2011, **3**, 33.

152 K. Gupta, D. R. Roy, V. Subramanian and P. K. Chattaraj, Are strong Brønsted acids necessarily strong Lewis acids?, *J. Mol. Struct.: THEOCHEM*, 2007, **812**, 13–24.

153 T. Koopmans, Über die Zuordnung von Wellenfunktionen und Eigenwerten zu den Einzelnen Elektronen Eines Atoms, *Physica*, 1934, **1**, 104–113.

154 L. R. Domingo, M. Ríos-Gutiérrez and P. Pérez, Applications of the Conceptual Density Functional Theory Indices to Organic Chemistry Reactivity, *Molecules*, 2016, **21**, 748.

155 P. Geerlings, F. De Proft and W. Langenaeker, Conceptual Density Functional Theory, *Chem. Rev.*, 2003, **103**, 1793–1874.

156 L. R. Domingo and P. Pérez, The nucleophilicity N index in organic chemistry, *Org. Biomol. Chem.*, 2011, **9**, 7168–7175.

157 L. R. Domingo, E. Chamorro and P. Pérez, Understanding the Reactivity of Captodative Ethylenes in Polar Cycloaddition Reactions. A Theoretical Study, *J. Org. Chem.*, 2008, **73**, 4615–4624.