Supporting Information for

OSCAR: An Extensive Repository of Chemically and Functionally Diverse Organocatalysts

Simone Gallarati,[a] Puck van Gerwen,[a,b] Ruben Laplaza,[a,b] Sergi Vela,[a] Alberto Fabrizio,[a,c] and Clemence Corminboeuf[a,b,c,*]

[a]Laboratory for Computational Molecular Design, Institute of Chemical Sciences and Engineering, Ecole Polytechnique Fédérale de Lausanne (EPFL), 1015 Lausanne, Switzerland

[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

*Email: clemence.corminboeuf@epfl.ch

Contents

1. Seed and CSD-Extracted Datasets

1.1 Function-based fragments

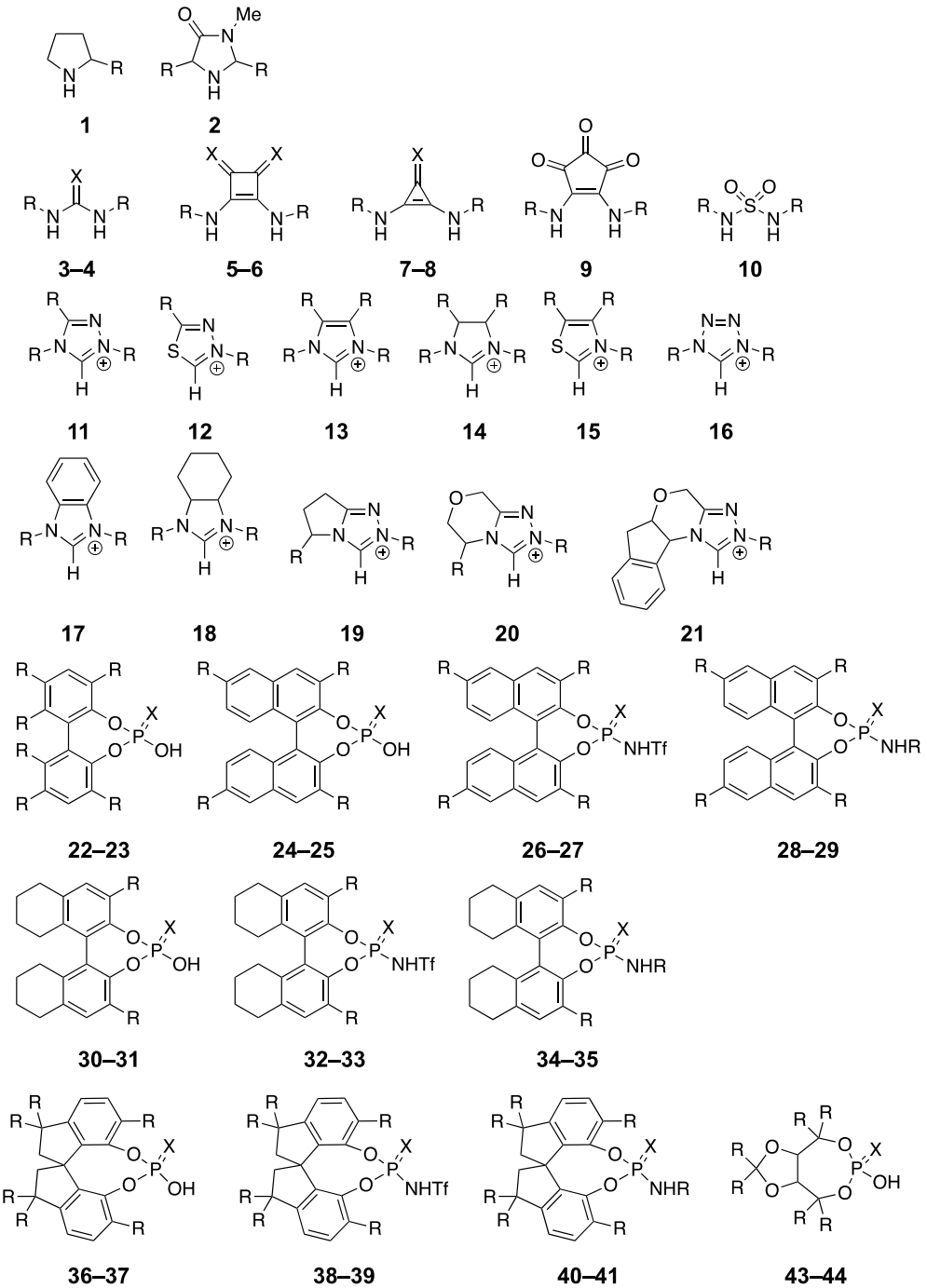

Figure S1. Function-based fragments searched in CSD (X= O/S).

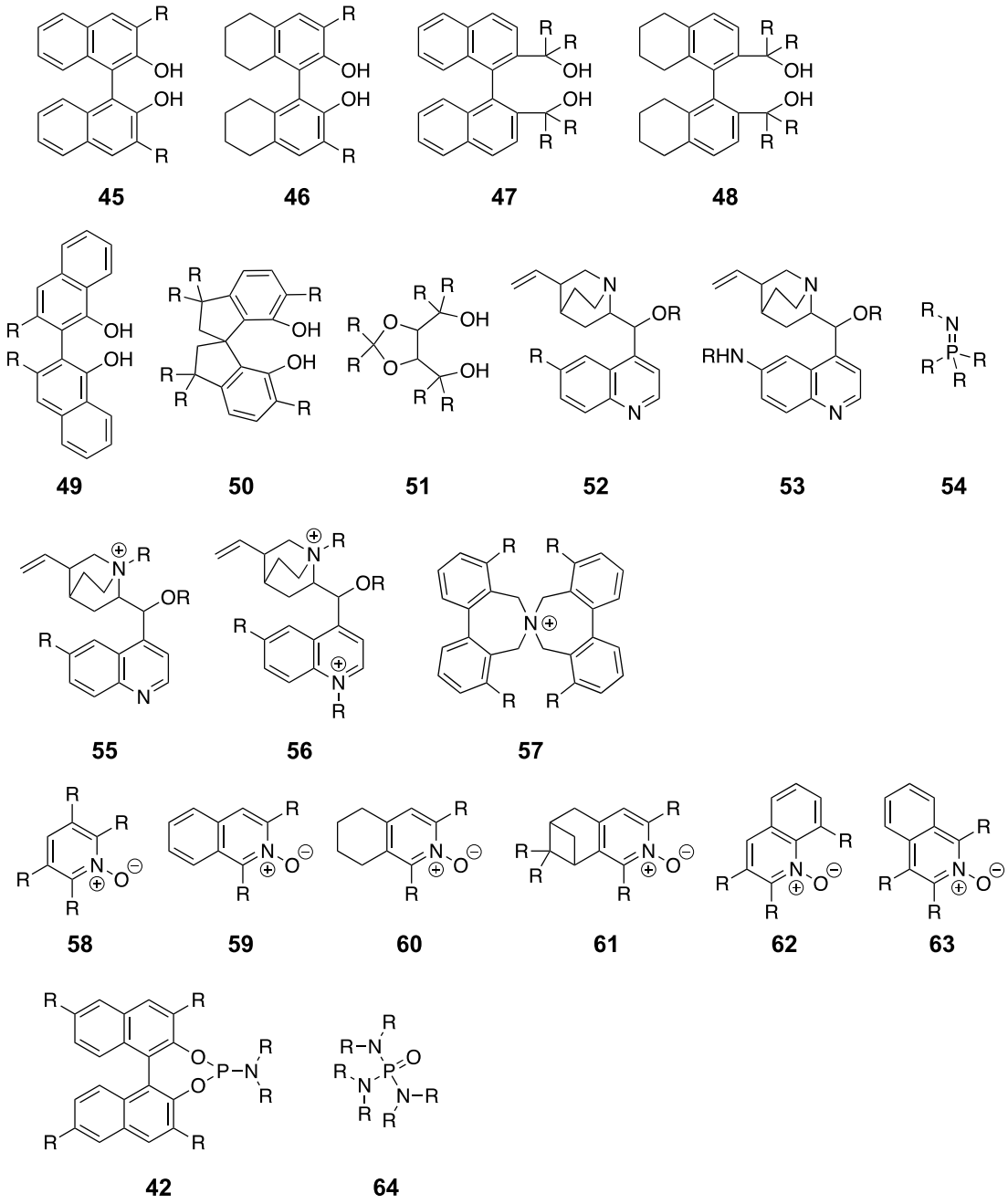

Figure S2. (Continued) Function-based fragments searched in CSD (X = O/S).

Table S1. SMILES strings of function-based fragments searched in CSD.

Motif	SMILES	Type
1	C1N([H])CCC1	Aminocat.
2	C1N([H])CC(N1C)=O	Aminocat.
3	N([H])C(N([H]))=O	DHBD
4	N([H])C(N([H]))=S	DHBD
5	O=C1C(N([H]))=C(N([H]))C1=O	DHBD
6	S=C1C(N([H]))=C(N([H]))C1=S	DHBD
7	O=C1C(N([H]))=C1N([H])	DHBD
8	S=C1C(N([H]))=C1N([H])	DHBD
9	O=C(C(N([H]))=C(N([H]))C1=O)C1=O	DHBD
10	O=S(N([H]))(N([H]))=O	DHBD
11	[H]C1=[N+]N=CN1	NHC
12	[H]C1=[N+]N=CS1	NHC
13	C1=C[N+]=C([H])N1	NHC
14	C1C[N+]=C([H])N1	NHC
15	C1=C[N+]=C([H])S1	NHC
16	[H]C1=[N+]N=NN1	NHC
17	N(C([H])=[N+]1)C2=C1C=CC=C2	NHC
18	N1C2CCCCC2[N+]=C1[H]	NHC
19	[H]C1=[N+]N=C2N1CCC2	NHC
20	[H]C1=[N+]N=C2N1CCOC2	NHC
21	[H]C1=[N+]N=C2N1C3C(CC4=C3C=CC=C4)OC2	NHC
22	C1=CC=CC(O2)=C1C3=CC=CC=C3OP2(O([H]))=O	PA
23	C1=CC=CC(O2)=C1C3=CC=CC=C3OP2(O([H]))=S	PA
24	O([H])P1(OC2=CC=C3C(C=CC=C3)=C2C4=C(O1)C=CC5=C4C=CC=C5)=O	PA
25	O([H])P1(OC2=CC=C3C(C=CC=C3)=C2C4=C(O1)C=CC5=C4C=CC=C5)=S	PA
26	O=P1(N([H])S(=O)(C(F)(F)F)=O)OC2=C(C3=C(C=CC=C4)C4=CC=C3O1)C5=C(C=CC=C5)C=C2	PA
27	S=P1(N([H])S(=O)(C(F)(F)F)=O)OC2=C(C3=C(C=CC=C4)C4=CC=C3O1)C5=C(C=CC=C5)C=C2	PA
28	O=P1(N([H]))OC2=C(C3=C(C=CC=C4)C4=CC=C3O1)C5=C(C=CC=C5)C=C2	PA
29	S=P1(N([H]))OC2=C(C3=C(C=CC=C4)C4=CC=C3O1)C5=C(C=CC=C5)C=C2	PA
30	O([H])P1(OC2=CC=C3C(CCCC3)=C2C4=C(O1)C=CC5=C4CCCC5)=O	PA
31	O([H])P1(OC2=CC=C3C(CCCC3)=C2C4=C(O1)C=CC5=C4CCCC5)=S	PA
32	O=P1(N([H])S(=O)(C(F)(F)F)=O)OC2=C(C3=C(CCCC4)C4=CC=C3O1)C5=C(CCCC5)C=C2	PA
33	S=P1(N([H])S(=O)(C(F)(F)F)=O)OC2=C(C3=C(CCCC4)C4=CC=C3O1)C5=C(CCCC5)C=C2	PA
34	O=P1(N([H]))OC2=C(C3=C(CCCC4)C4=CC=C3O1)C5=C(CCCC5)C=C2	PA
35	S=P1(N([H]))OC2=C(C3=C(CCCC4)C4=CC=C3O1)C5=C(CCCC5)C=C2	PA
36	O=P1(O([H]))OC2=CC=CC3=C2C4(CCC5=C4C(O1)=CC=C5)CC3	PA
37	S=P1(O([H]))OC2=CC=CC3=C2C4(CCC5=C4C(O1)=CC=C5)CC3	PA
38	O=P1(N([H])S(=O)(C(F)(F)F)=O)OC2=CC=CC3=C2C4(CCC5=C4C(O1)=CC=C5)CC3	PA
39	S=P1(N([H])S(=O)(C(F)(F)F)=O)OC2=CC=CC3=C2C4(CCC5=C4C(O1)=CC=C5)CC3	PA
40	O=P1(N([H]))OC2=CC=CC3=C2C4(CCC5=C4C(O1)=CC=C5)CC3	PA
41	S=P1(N([H]))OC2=CC=CC3=C2C4(CCC5=C4C(O1)=CC=C5)CC3	PA
42	C1=CC(C=CC=C2)=C2C3=C1OP(N)OC4=CC=C5C(C=CC=C5)=C34	LB
43	C(OP(OC1)(O([H]))=O)C2C1OCO2	PA
44	C(OP(OC1)(O([H]))=S)C2C1OCO2	PA
45	O([H])C1=CC=C2C(C=CC=C2)=C1C3=C(O([H]))C=CC4=C3C=CC=C4	SHBD
46	O([H])C1=CC=C2C(CCCC2)=C1C3=C(O([H]))C=CC4=C3CCCC4	SHBD
47	O([H])CC1=C(C2=C(C=CC=C3)C3=CC=C2C(O([H])))C4=C(C=CC=C4)C=C1	SHBD
48	O([H])CC1=C(C2=C(CCCC3)C3=CC=C2C(O([H])))C4=C(CCCC4)C=C1	SHBD
49	O([H])C1=C2C(C=CC=C2)=CC=C1C3=C(O([H]))C(C=CC=C4)=C4C=C3	SHBD
50	O([H])C1=CC=CC2=C1C3(CC2)CCC4=C3C(O([H]))=CC=C4	SHBD
51	O([H])CC1C(CO([H]))OCO1	SHBD
52	C=CC(C(CC1)C2)CN1C2C(O)C3=CC=NC4=C3C=CC=C4	BB
53	C=CC(C(CC1)C2)CN1C2C(O)C3=CC=NC4=C3C=C(N([H]))C=C4	BB
54	P=N	BB
55	C=CC(C(CC1)C2)C[N+]1C2C(O)C3=CC=NC4=C3C=CC=C4	LA
56	C=CC(C(CC1)C2)C[N+]1C2C(O)C3=CC=[N+]C4=C3C=CC=C4	LA
57	C1(C2=CC=CC=C2C[N+]3(CC(C=CC=C4)=C4C(C=CC=C5)=C5C3)C6)=C6C=CC=C1	LA
58	C1=[N+]([O-])C=CC=C1	LB
59	C1=[N+]([O-])C=CC2=C1C=CC=C2	LB
60	C1=[N+]([O-])C=CC2=C1CCCC2	LB
61	C1=[N+]([O-])C=CC2=C1C3CC(C3)C2	LB
62	C1=[N+]([O-])C(C=CC=C2)=C2C=C1	LB
63	C1=[N+]([O-])C=C(C=CC=C2)C2=C1	LB
64	O=P(N)(N)N	LB

1.2 Composition of datasets

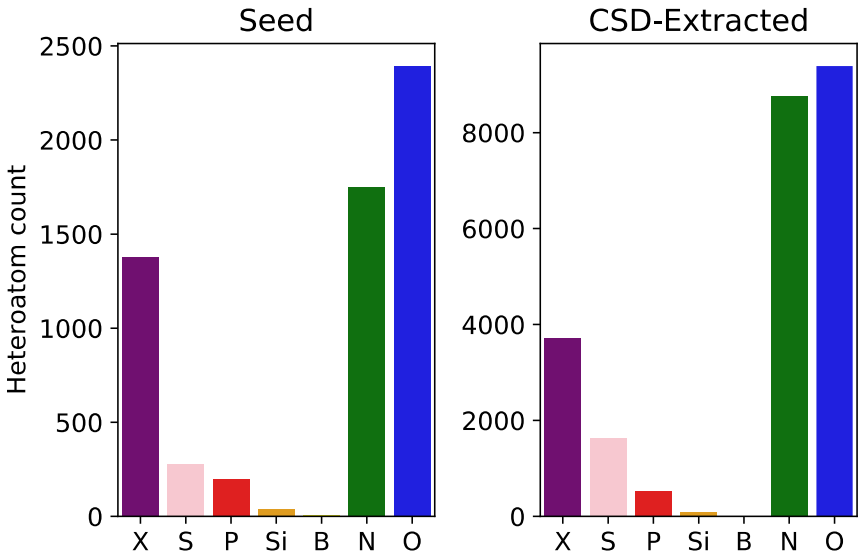

Figure S3. Distribution histograms of heteroatom types (X = halogen) in the seed and CSD-extracted datasets.

Table S2. Count of functional groups (FG) present in the seed and CSD-extracted datasets.

Atom type	FG	Description	Seed	CSD
C	$R_3C–CR_3$	Alkane	5950	9553
C	$R_2C=CR_2$	Alkene	134	365
C	RC#CR	Alkyne	4	34
C	$R_2C=C=CR_2$	Allene	2	0
N	NR_3	Amine	993	1365
N	C=N	Imine	574	1663
N	C(=O)N	Amide	238	2600
N	$R_2C=CNR_2$	Enamine	68	111
N	NR_4^+	Ammonium	52	43
N	$R_2C=NR_2^+$	Iminium	29	1
N	N=N	Azo	14	19
N	N=N=N	Azide	11	17
N	N–B	Amine-borane	4	0
N	RC#N	Nitrile	3	133
N	N–N	Hydrazine	0	14
O	C–O–C	Ether	578	774
O	C(C=O)C	Ketone	219	311
O	RC(=O)OR	Carboxylic	175	1009
O	C=N(O)–R	Nitrone	136	642
O	RNO_2	Nitro	28	232
O	N–O	Other	4	26
O	$ROCO_2R$	Carbonate	0	6
O	RO–OR	Peroxide	0	3
O	RO^-	Alkoxide	0	3
O	$RCOO^-$	Carboxylate	0	2
O	RN=O	Nitroso	0	1
O	N=C=O	Isocyanate	0	1
S	CSN	Thioamide	230	1905
S	S–N	Thioamine	94	112
S	SO_2R_2	Sulfone	93	122

S	C–S–C	Thioether	62	172
S	C–CS–C	Thiocarbonyl	28	4
S	SO_3R	Sulfonate	13	14
S	S–P	Thiophosphine	5	5
S	SOR_2	Sulfoxide	5	6
S	CS_2R	Dithiocarboxylic	1	3
S	COSR	Thiocarboxylic	0	2
S	SO_2R	Sulfinate	0	1
S	CSOR	Thiocarboxylic	0	1
P	PR_3	Phosphine	68	12
P	PO_4	Phosphate	64	13
P	P=C	Ylide	47	459
P	POR_3	Phosphine oxide	30	45
P	PO_3N	Phosphoramidate	27	12
P	P–O	Phosphoxide	12	5
P	P–N	Phosphinamine	6	6
P	P=N	Phosphazene	2	170
P	PO_3R	Phosphonate	1	21
P	PO_2N	Phosphoramidite	1	33
X	C–X	Carbon halide	1611	2034
X	F	Fluorine	1557	1200
X	Cl	Chlorine	47	595
X	Br	Bromine	19	240
X	S–X	Sulfur halide	10	0
X	B–X	Boron halide	2	0
X	N–X	Nitrogen halide	0	1
B	BOR_2	Borninate	4	3
B	BR_3	Borane	2	0
B	BO_2R	Boronate	2	0
Ring	6-mem-ring		19917	36926
Ring	5-mem-ring		3051	8870
Ring	7-mem-ring		520	243
Ring	4-mem-ring		300	248
Ring	3-mem-ring		24	92

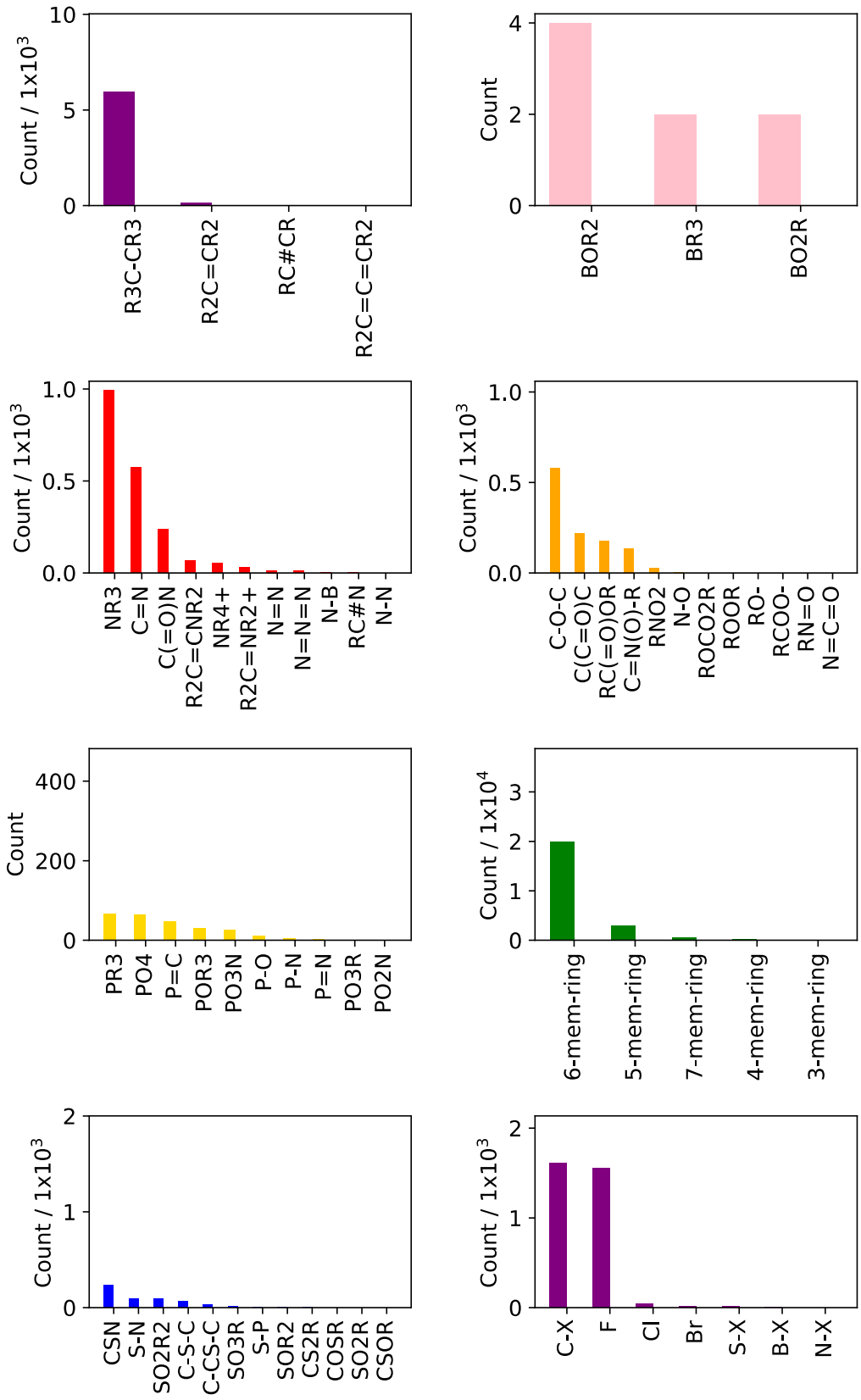

Figure S4. Bar plots of functional groups in the seed (left-hand-side, darker colors) and CSD-extracted (right-hand-side, lighter colors) datasets (data from Table S2).

1.3 *cell2mol* usage

cell2mol[1] is freely available at https://github.com/lcmd-epfl/cell2mol. Detailed information regarding installation, usage, and a worked example is provided on GitHub. Herein, *cell2mol* was used to interpret molecular crystals and retrieve information about the individual molecules, including the number of atoms (N), the total molecular charge (Q), the number and type of elements (*i.e.*, the formula), and the number and type of adjacencies (defined as occurrences of neighboring pairs of elements) of selected CSD entries that were identified by searching the function-based fragments in section 1.1 with ConQuest. All the above variables are stored during the *cell2mol* analysis and were used to detect and eliminate duplicate entries from OSCAR: if two or more molecules share the same N, Q, formula, and adjacencies, they are considered equal and only one is retained.

1.4 Alternative structure maps

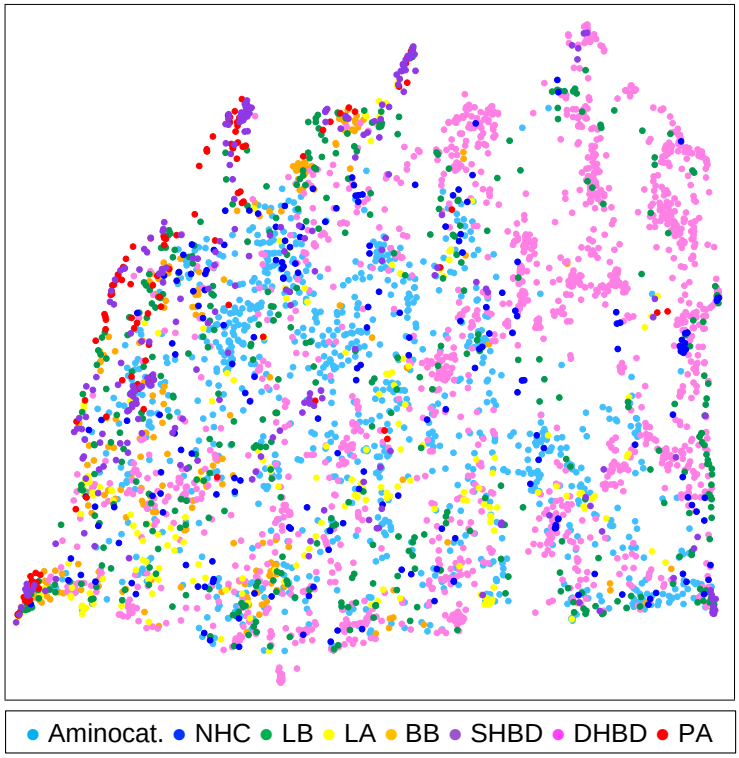

Figure S5. 2D UMAP map of OSCAR (seed + CSD-extracted datasets) on the basis of the organocatalysts' FCHL19 representation.

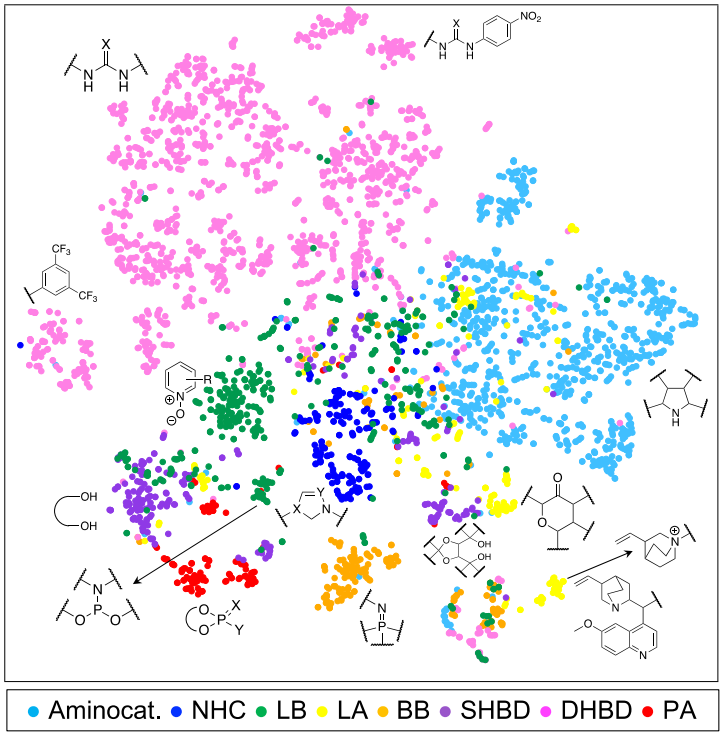

Figure S6. 2D t-SNE map of OSCAR (seed + CSD-extracted datasets) on the basis of the organocatalysts' Morgan fingerprints.

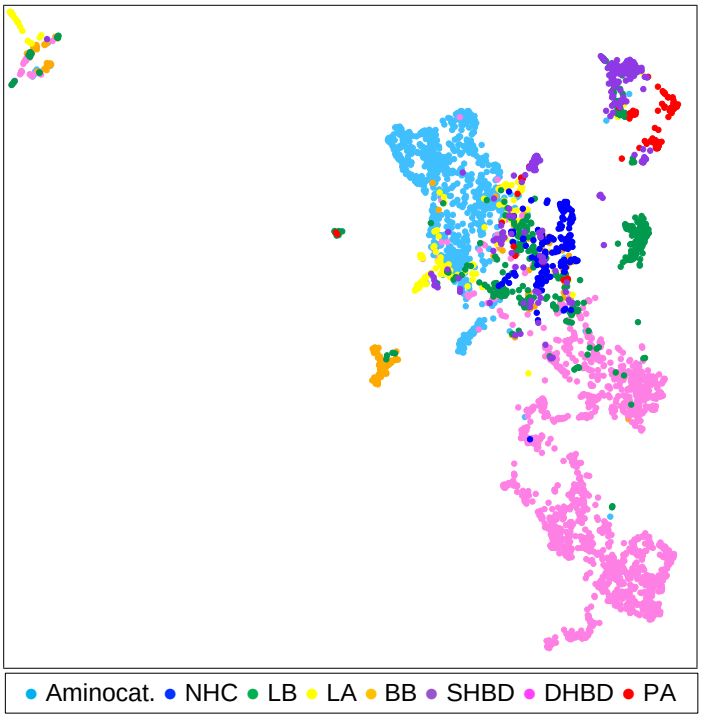

Figure S7. 2D UMAP map of OSCAR (seed + CSD-extracted datasets) on the basis of the organocatalysts' Morgan fingerprints.

FCHL19 fingerprints[2] were generated with the QML Python package[3] and used as input for t-SNE[4] and UMAP[5] dimensionality reduction algorithms (Figure 4A and Figure S5, respectively). Morgan fingerprints (also known as extended connectivity fingerprints, ECFP)[6] with radius 2 were generated with RDKit[7] (version 2020.09.1.0) and used as molecular representation for dimensionality reduction (Figure S6 and Figure S7). Since the initial feature dimensions are high (3,408 and 512 for FCHL19 and ECFP, respectively), truncated SVD decomposition[8] was first applied using the Python package Scikit-learn (version 0.23.2) to reduce the dimensionality to 50. t-SNE maps[4] were then generated, reducing the dimensions from 50 to 2. The perplexity parameter was set to 20, and all other parameters left as default. UMAP maps[5] were generated using the UMAP-learn Python package (version 0.5.3) with 50 nearest neighbors.

The FCHL19 representation (a non-linear potential exploiting atom types and positions) was chosen to generate the 2D t-SNE map in Figure 4A for its ability to capture both the chemistry of the organocatalysts' catalytic motifs and functional groups, as well as their 3D atomic arrangements. Together with physics-based features (*e.g.*, FCHL19), the t-SNE algorithm well preserves the local similarity between molecules.[9] Morgan fingerprints successfully cluster organocatalysts according to their catalytic motif, however they carry no information regarding their spatial arrangements. Both t-SNE and UMAP maps generated using Morgan fingerprints (Figure S6 and Figure S7) were successful at clustering similar species according to their chemical functionalities, however UMAP (Figure S7) specifically isolated a small group of *cinchona* derivatives (top left cluster), with worse space separation of the other families of catalysts. The t-SNE and UMAP algorithms vary in the manner in which they optimize the low-dimensional manifolds, with UMAP using cross-entropy rather than the Kullback-Leibler (KL) divergence. Cross-entropy loss better preserves *global* structure, whereas minimizing the KL divergence better preserves *local* structure. In other words, while points in each cluster in a t-SNE map are guaranteed to be similar, there is not a reliable notion of similarity between clusters. The small cluster of points (bottom right corner) in the t-SNE map (Figure S6) therefore looks less dissimilar to the other clusters than it does in the UMAP map (Figure S7).

2. OSCAR!(NHC)

2.1 Stereoelectronic descriptors

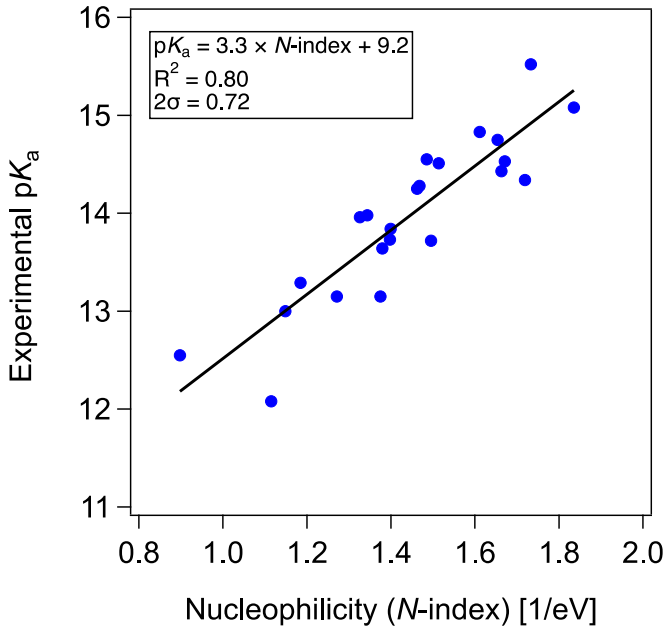

Figure S8. Experimental pK_a of azolium ion precursors *vs*. nucleophilicity (*N*-index, ωB97X-D/Def2-TZVP//B97-D/Def2-TZVP) of selected carbenes.[10]

Table S3. Experimental pKa's of azolium ions vs. nucleophilicity (*N*-index, ωB97X-D/Def2-TZVP//B97-D/Def2-TZVP) of selected carbenes from the literature.[10]

Structure	SMILES	pK_a	*N*-index (1/eV)
carbene_100	C12=NN(C3=CC=CC=C3)[C]N1[C@H]4[C@H](CC5=C4C=CC=C5)OC2	13.15	1.37
carbene_101	CC(C=C(C)C=C1C)=C1N2[C]N3C(CO[C@@H]4[C@H]3C5=C(C=CC=C5)C4)=N2	13.72	1.50
carbene_102	CC(C=C(C)C=C1C)=C1N2C(C3=CC=CC=C3)=NN(C4=CC=CC=C4)[C]2	13.00	1.15
carbene_103	CC(C=C(C)C=C1C)=C1N2C(C3=CC=CC=C3)=NN(C4=C(C)C=C(C)C=C4C)[C]2	13.29	1.18
carbene_45	CC1=CC(C)=CC(C)=C1N2[C]N3C(CCC3)=N2	15.52	1.73
carbene_46	C12=NN(C3=CC=CC=C3)[C]N1CCC2	14.55	1.48
carbene_83	COC(C=C1)=CC=C1N2[C]N3C(CCC3)=N2	15.08	1.84
carbene_84	CC(C=C1)=CC=C1N2[C]N3C(CCC3)=N2	14.83	1.61
carbene_85	FC(C=C1)=CC=C1N2[C]N3C(CCC3)=N2	14.28	1.47
carbene_86	ClC(C=C1)=CC=C1N2[C]N3C(CCC3)=N2	13.98	1.34
carbene_87	BrC(C=C1)=CC=C1N2[C]N3C(CCC3)=N2	13.96	1.33
carbene_88	FC(C(F)=C1F)=C(F)C(F)=C1N2[C]N3C(CCC3)=N2	12.08	1.12
carbene_89	C12=NN(C3=CC=CC=C3)[C]N1[C@H](C(C4=CC=CC=C4)C5=CC=CC=C5)CC2	14.25	1.46
carbene_90	C12=NN(C3=CC=CC=C3)[C]N1CCOC2	13.84	1.40
carbene_91	COC(C=C1)=CC=C1N2[C]N3C(COCC3)=N2	14.34	1.72
carbene_92	FC(C=C1)=CC=C1N2[C]N3C(COCC3)=N2	13.64	1.38
carbene_93	ClC(C=C1)=CC=C1N2[C]N3C(COCC3)=N2	13.15	1.27
carbene_94	N#CC(C=C1)=CC=C1N2[C]N3C(COCC3)=N2	12.55	0.90
carbene_95	CC(C=C1C)=CC(C)=C1N2[C]N3C(COCC3)=N2	14.75	1.65
carbene_96	C[C@@H](CC)[C@H]1COCC2=NN(C3=CC=CC=C3)[C]N21	13.73	1.40
carbene_97	CC(C)(C)[C@H]1COCC2=NN(C3=C(C)C=C(C)C=C3C)[C]N21	14.43	1.66
carbene_98	CC(C)[C@H]1COCC2=NN(C3=C(C)C=C(C)C=C3C)[C]N21	14.53	1.67
carbene_99	CC(C=C(C)C=C1C)=C1N2[C]N3C(COC[C@@H]3CC4=CC=CC=C4)=N2	14.51	1.51

The percentage buried volume (%V_{buried}) was computed with the *MolVol* package, which is freely available at https://github.com/lcmd-epfl/molvol using a standard 3.5 Å sphere placed 2.0 Å away from

the carbene in the direction given by the bisection of the N–C–N angle. For all atoms, Bondi radii scaled by 1.17 (as in the SambVca implementation)[11] were used.

2.2 Structures generation

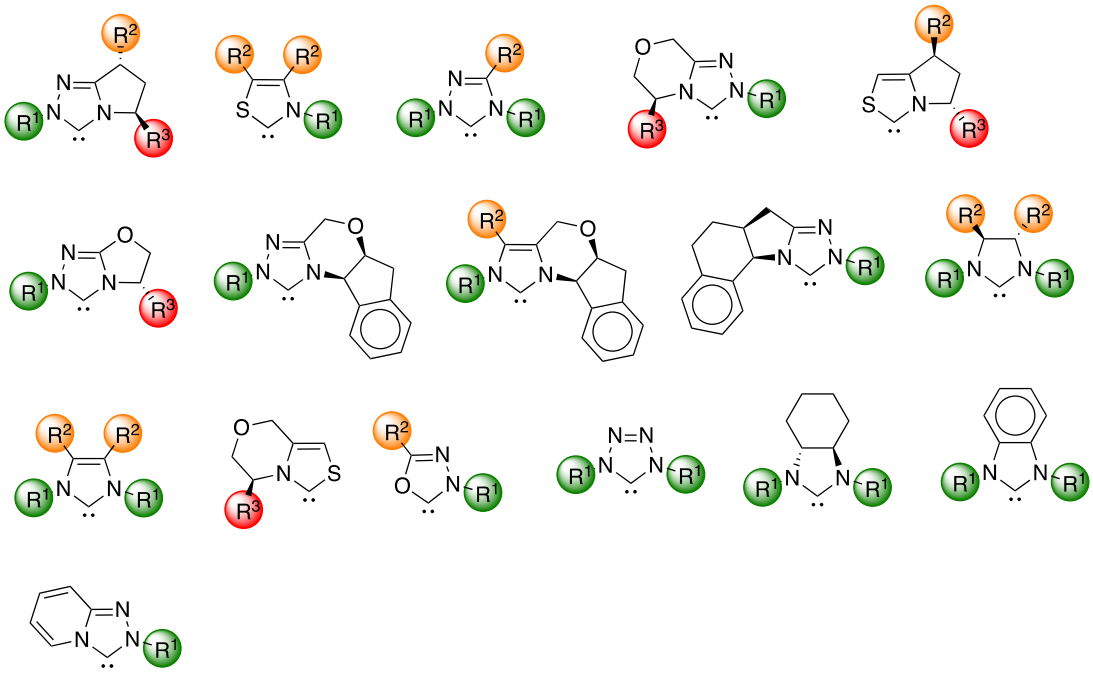

Figure S9. Scaffolds used to generate combinatorial structures in OSCAR!(NHC).

Table S4. List of flexible SMILES strings of carbene scaffolds for OSCAR!(NHC).

Scaffold	Flexible SMILES
1	[C@H]1(C[C@H](N2C1=NN([C]2)R^1)R^3)R^2
2	C1(=C(N([C]S1)R^1)R^2)R^2
3	N1=C(N([C]N1(R^1))R^1)R^2
4	C1([C@@H](N2[C]N(N=C2CO1)R^1)R^3)([H])([H])
5	[C@@H]1(C[C@@H](N2C1=CS[C]2)R^3)R^2
6	[C@@H]1(COC2=NN([C]N12)R^1)R^3
7	[C@H]12Cc3ccccc3[C@H]1N4[C]N(N=C4CO2)R^1
8	C1(=C2CO[C@H]3Cc4ccccc4[C@H]3N2[C]N1(R^1))R^2
9	[C@H]12CC3=NN([C]N3[C@H]1c5ccccc5CC2)R^1
10	[C@H]1(N([C]N([C@H]1R^2)R^1)R^1)R^2
11	C1(=C(N([C]N1R^1)R^1)R^2)R^2
12	C1([C@@H](N2[C]S(N=C2CO1))R^3)([H])([H])
13	C1(=NN([C]O1)R^1)R^2
14	N1([C]N(N=N1)R^1)R^1
15	N1([C]N([C@H]4[C@H]1CCCC4)R^1)R^1
16	N1([C]N(c4ccccc14)R^1)R^1
17	N12[C]N(N=C1C=CC=C2)R^1

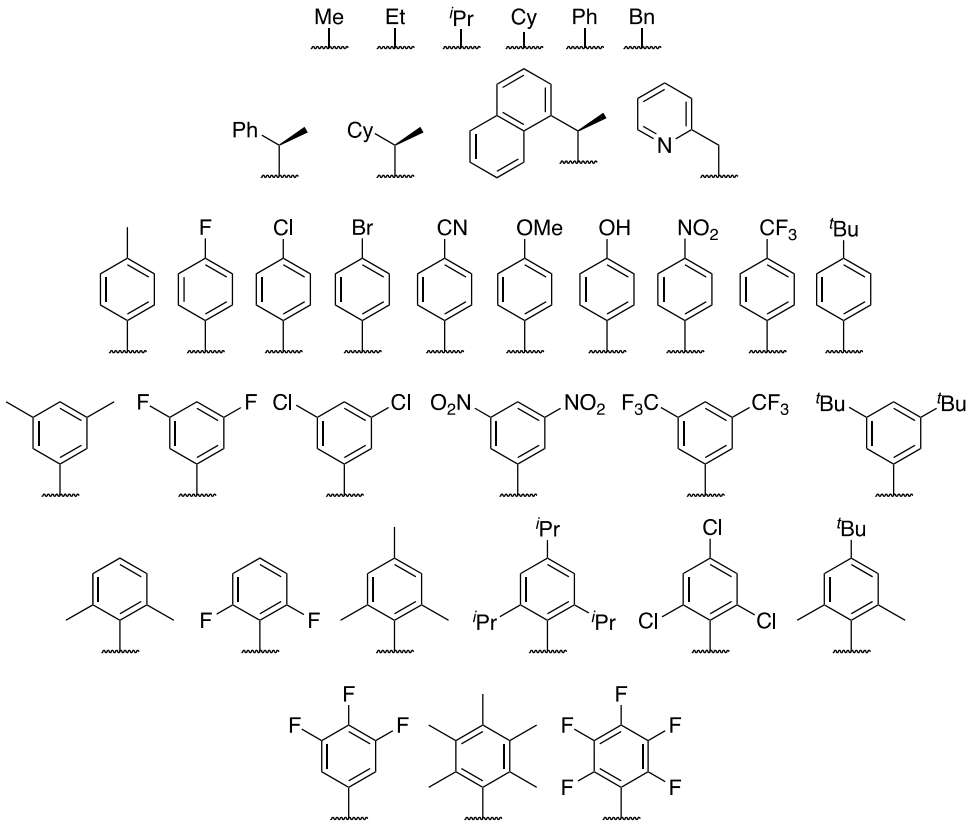

Figure S10. List of R^1 substituents for OSCAR!(NHC).

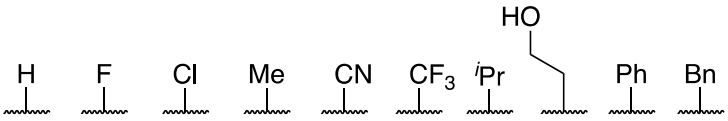

Figure S11. List of R^2 substituents for OSCAR!(NHC).

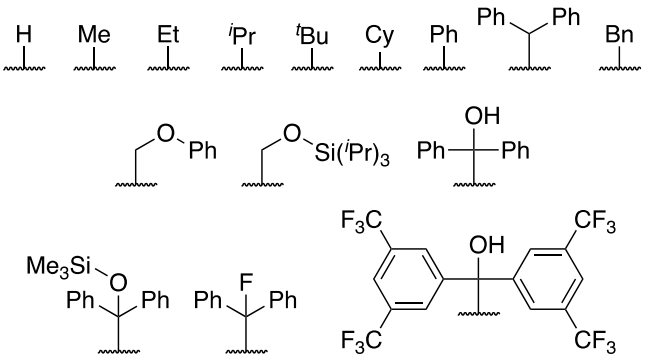

Figure S12. List of R^3 substituents for OSCAR!(NHC).

Table S5. List of SMILES strings of $R^{1–3}$ substituents for OSCAR!(NHC).

Substituent #	R	SMILES
1	R^1	C
2	R^1	CC
3	R^1	C(C)(C)
4	R^1	C1CCCCC1
5	R^1	C1=CC=CC=C1
6	R^1	CC1=CC=CC=C1
7	R^1	[C@H](C)C1=CC=CC=C1
8	R^1	[C@H](C)C1CCCCC1
9	R^1	[C@H](C)C1=CC=CC2=C1C=CC=C2
10	R^1	CC1=NC=CC=C1
11	R^1	C1=CC=C(C)C=C1
12	R^1	C1=CC=C(F)C=C1
13	R^1	C1=CC=C(Cl)C=C1
14	R^1	C1=CC=C(Br)C=C1
15	R^1	C1=CC=C(C#N)C=C1
16	R^1	C1=CC=C(OC)C=C1
17	R^1	C1=CC=C(O)C=C1
18	R^1	C1=CC=C([N+](=O)([O-]))C=C1
19	R^1	C1=CC=C(C(F)(F)(F))C=C1
20	R^1	C1=CC=C(C(C)(C)(C))C=C1
21	R^1	C1=CC(C)=CC(C)=C1
22	R^1	C1=CC(F)=CC(F)=C1
23	R^1	C1=CC(Cl)=CC(Cl)=C1
24	R^1	C1=CC([N+]([O-])(=O))=CC([N+]([O-])(=O))=C1
25	R^1	C1=CC(C(F)(F)(F))=CC(C(F)(F)(F))=C1
26	R^1	C1=CC(C(C)(C)(C))=CC(C(C)(C)(C))=C1
27	R^1	C1=C(C)C=CC=C1(C)
28	R^1	C1=C(C)C=C(C)C=C1(C)
29	R^1	C1=C(C(C)(C))C=C(C(C)(C))C=C1(C(C)(C))
30	R^1	C1=C(Cl)C=C(Cl)C=C1(Cl)
31	R^1	C1=C(C)C=C(C(C)(C)(C))C=C1(C)
32	R^1	C1=CC(F)=C(F)C(F)=C1
33	R^1	C1=C(F)C=CC=C1(F)
34	R^1	C1=C(C)C(C)=C(C)C(C)=C1(C)
35	R^1	C1=C(F)C(F)=C(F)C(F)=C1(F)
1	R^2	[H]
2	R^2	F
3	R^2	Cl
4	R^2	C
5	R^2	C#N
6	R^2	C(F)(F)(F)
7	R^2	C(C)(C)
8	R^2	CCO
9	R^2	C1=CC=CC=C1
10	R^2	CC1=CC=CC=C1
1	R^3	[H]
2	R^3	C
3	R^3	CC
4	R^3	C(C)(C)
5	R^3	C(C)(C)(C)
6	R^3	C1CCCCC1
7	R^3	C1=CC=CC=C1
8	R^3	C(C4=CC=CC=C4)C5=CC=CC=C5
9	R^3	CC1=CC=CC=C1
10	R^3	COC1=CC=CC=C1
11	R^3	CO[Si](C(C)C)(C(C)C)(C(C)C)
12	R^3	C(C4=CC=CC=C4)(C5=CC=CC=C5)O
13	R^3	C(C4=CC=CC=C4)(C5=CC=CC=C5)O[Si](C)(C)(C)
14	R^3	C(C4=CC=CC=C4)(C5=CC=CC=C5)F
15	R^3	C(C4=CC(C(F)(F)(F))=CC(C(F)(F)(F))=C4)(C5=CC(C(F)(F)(F))=CC(C(F)(F)(F))=C5)O

3. OSCAR!(DHBD)

3.1 Electronic descriptor

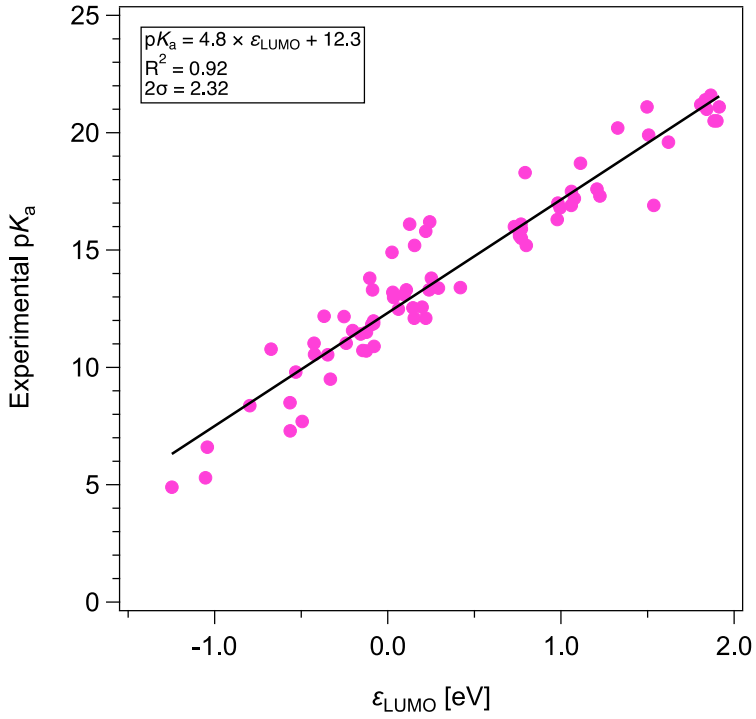

Figure S13. Experimental pK_a *vs.* ε_{LUMO} (ωB97X-D/Def2-TZVP//B97-D/Def2-TZVP) of selected DHBDs.[12]

Table S6. Experimental pK_a's and ε_{LUMO} (ωB97X-D/Def2-TZVP//B97-D/Def2-TZVP) of selected DHBDs from the literature.[12]

Structure	SMILES	pK_a	ε_{LUMO} (eV)
Ur2	O=C(NC1=CC=CC=C1)NC2=CC=CC=C2	18.7	1.11
Ur3	O=C(NC1=CC(C(F)(F)F)=CC(C(F)(F)F)=C1)NC2=CC=CC=C2	16.1	0.13
Ur4	O=C(NC1=CC(C(F)(F)F)=CC(C(F)(F)F)=C1)NC2=CC(C(F)(F)F)=CC(C(F)(F)F)=C2	13.8	-0.10
Ur6	O=C(NC1=CC=CC(C(F)(F)F)=C1)NC2=CC(C(F)(F)F)=CC(C(F)(F)F)=C2	14.9	0.02
ThUr1	S=C(N[C@H]1[C@H](N(C)C)CCCC1)NC2=CC(C(F)(F)F)=CC(C(F)(F)F)=C2	13.8	0.25
ThUr2	NC(N)=S	21.1	1.50
ThUr3	S=C(NC1=CC=CC=C1)NC2=CC=CC=C2	13.4	0.42
ThUr4	S=C(NC1=CC=CC(C(F)(F)F)=C1)NC2=CC=CC=C2	12.1	0.15
ThUr5	S=C(NC1=CC(C(F)(F)F)=CC(C(F)(F)F)=C1)NC2=CC=CC=C2	10.7	-0.13
ThUr6	S=C(N[C@H](C1=C(C=C(OC)C=C2)C2=NC=C1)[C@@H]3[N@](CC4)C[C@@H](C=C)[C@H]4C3)NC5=CC(C(F)(F)F)=CC(C(F)(F)F)=C5	13.2	0.03
ThUr7	S=C(NC1=CC=CC(C(F)(F)F)=C1)NC2=CC(C(F)(F)F)=CC=C2	10.9	-0.08
ThUr8	S=C(NC1=CC(C(F)(F)F)=CC(C(F)(F)F)=C1)NC2=CC(C(F)(F)F)=CC(C(F)(F)F)=C2	8.5	-0.57
ThUr9	S=C(N[C@H]1[C@H](N2C(C)=CC=C2C)CCCC1)N[C@@H](C(C)(C)C)C(N(CC)CC)=O	19.6	1.62
ThUr10	S=C(NC1=CC(C(F)(F)F)=CC(C(F)(F)F)=C1)NC2=C(C3=C(C=CC=C4)C4=CC=C3N(C)C)C5=CC=CC=C5C=C2	10.72	-0.15
ThUr11	S=C(NC1=CC(C(F)(F)F)=CC(C(F)(F)F)=C1)N[C@@H]2CCCC[C@H]2NC(NC3=CC(C(F)(F)F)=CC(C(F)(F)F)=C3)=S	11.98	-0.08
ThUr12	S=C(N[C@H]1[C@H](N(C)C)CCCC1)NC	20.5	1.88
ThUr13	S=C(N[C@H]1[C@H](N(C)C)CCCC1)NCC	21	1.84
ThUr14	S=C(N[C@H]1[C@H](N(C)C)CCCC1)NC(C)C	21.6	1.87
ThUr15	S=C(N[C@H]1[C@H](N(C)C)CCCC1)NCC2=CC=CC=C2	20.2	1.33
ThUr16	S=C(N[C@H]1[C@H](N(C)C)CCCC1)NC2CCCCC2	20.5	1.90
ThUr17	S=C(N[C@H]1[C@H](N(C)C)CCCC1)NC2=CC=CC=C2	17	0.98
ThUr18	S=C(N[C@H]1[C@H](N(C)C)CCCC1)NC(C)(C)C	21.4	1.83
ThUr19	S=C(N[C@H]1[C@H](N(C)C)CCCC1)NC(CC)(CC)CC	21.2	1.81
ThUr20	S=C(NC1=CC(C(F)(F)F)=CC(C(F)(F)F)=C1)N[C@H]2[C@H](N3C(C4=CC=CC=C4)=CN=C3)CCCC2	12.54	0.14
ThUr21	O=C([C@@H](NC(NC1=CC(C(F)(F)F)=CC(C(F)(F)F)=C1)=S)C(C)(C)C)N(C)CC2=CC=CC=C2	12.57	0.20
ThUr22	S=C(N[C@H]1[C@@H](O)CC2=C1C=CC=C2)NC3=CC(C(F)(F)F)=CC(C(F)(F)F)=C3	12.98	0.03
ThUr23	S=C(NC[C@H]1NCCC1)NC2=CC(C(F)(F)F)=CC(C(F)(F)F)=C2	13.38	0.29
ThUr24	O=C([C@H](C(C)(C)C)NC(N[C@H]1[C@H](/N=C/C2=C(O)C(C(C)(C)C)=CC(C(C)(C)C)=C2)CCCC1)=S)N(CC)CC	18.3	0.79
ThUr25	S=C(N[C@@H](CC1=CC=CC=C1)CN(CC)CC)NC2=CC(C(F)(F)F)=CC(C(F)(F)F)=C2	13.3	0.24
ThUr26	S=C(N[C@H](C1=C(C=C(OC)C=C2)C2=NC=C1)[C@@H]3[N@](CC4)C[C@@H](C=C)[C@H]4C3)NC5=CC=CC=C5	15.8	0.22
ThUr27	S=C(N[C@H]1[C@H](N(C)C)CCCC1)NC2=CC=C(OC)C=C2	17.6	1.21
ThUr28	S=C(N[C@H]1[C@H](N(C)C)CCCC1)NC2=CC=C(C)C=C2	17.5	1.06
ThUr29	S=C(N[C@H]1[C@H](N(C)C)CCCC1)NC2=CC=C(Cl)C=C2	16.1	0.77
ThUr30	S=C(N[C@H]1[C@H](N(C)C)CCCC1)NC2=CC(Cl)=CC=C2	15.7	0.76
ThUr31	S=C(N[C@H]1[C@H](N(C)C)CCCC1)NC2=CC=CC=C2Cl	15.2	0.80
ThUr32	S=C(N[C@H]1[C@H](N(C)C)CCCC1)NC2=CC=C(Br)C=C2	16	0.73
ThUr33	S=C(N[C@H]1[C@H](N(C)C)CCCC1)NCCCC	21.1	1.91
ThUr34	S=C(N[C@H](C1=C(C=C(OC)C=C2)C2=NC=C1)[C@@H]3[N@](CC4)C[C@@H](C=C)[C@H]4C3)NC5=CC=C(C)C=C5	16.2	0.24
ThUr35	S=C(N[C@H](C1=C(C=C(OC)C=C2)C2=NC=C1)[C@@H]3[N@](CC4)C[C@@H](C=C)[C@H]4C3)NC5=CC=C(Cl)C=C5	15.2	0.15
ThUr37	S=C(NC1=CC=C(C)C=C1)N[C@@H](CC2=CC=CC=C2)CN(CC)CC	16.9	1.06
ThUr38	S=C(NC1=CC=CC=C1)N[C@@H](CC2=CC=CC=C2)CN(CC)CC	16.3	0.98
ThUr39	S=C(NC1=CC=C(Cl)C=C1)N[C@@H](CC2=CC=CC=C2)CN(CC)CC	15.5	0.77
ThUr40	S=C(NCCCC)N[C@@H](CC1=CC=CC=C1)CN(CC)CC	19.9	1.51
ThUr41	S=C(NC1=CC=C(OC)C=C1)NC2CCCCC2	17.3	1.22
ThUr42	S=C(NC1=CC=C(C)C=C1)NC2CCCCC2	17.2	1.08
ThUr43	S=C(NC1=CC=CC=C1)NC2CCCCC2	16.8	0.99
ThUr44	S=C(NC1=CC=C(Cl)C=C1)NC2CCCCC2	15.9	0.77
ThUr45	S=C(NC1=CC=CC(Cl)=C1)NC2CCCCC2	15.6	0.76
ThUr46	CN(C)[C@H]1CCCC[C@@H]1NC(NCC(F)(F)F)=S	16.9	1.54
ThUr47	S=C(NC1=CC=CC(C(F)(F)F)=C1)NC2=CC(C(F)(F)F)=CC(C(F)(F)F)=C2	9.5	-0.33
Sq1	O=C1C(NC2=CC(C(F)(F)F)=CC(C(F)(F)F)=C2)=C(N[C@H]3[C@H](N(C)C)CCCC3)C1=O	11.83	-0.09
Sq2	O=C1C(NC2=CC(C(F)(F)F)=CC(C(F)(F)F)=C2)=C(N[C@H]3[C@H](N4CCCC4)CCCC3)C1=O	13.3	-0.09
Sq3	O=C1C(NC2=CC=CC=C2)=C(NC3=CC=CC=C3)C1=O	12.48	0.06
Sq4	O=C1C(NC2=CC(C(F)(F)F)=CC(C(F)(F)F)=C2)=C(NC3=CC=CC=C3)C1=O	10.55	-0.42

Sq5	O=C1C(NC2=CC(C(F)(F)F)=CC(C(F)(F)F)=C2)=C(N[C@@H]([C@@H]3[N@](CC4)C[C@@H](C=C)[C@H]4C3)C5=C(C=C(OC)C=C6)C6=NC=C5)C1=O	10.54	-0.35
Sq6	O=C1C(NC2=CC(C(F)(F)F)=CC(C(F)(F)F)=C2)=C(NC3=CC(C(F)(F)F)=CC(C(F)(F)F)=C3)C1=O	8.37	-0.80
Sq7	O=C1C(NC2=CC=C(C(F)(F)F)C=C2)=C(N[C@@H]([C@@H]3[N@](CC4)C[C@@H](C=C)[C@H]4C3)C5=C(C=C(OC)C=C6)C6=NC=C5)C1=O	12.17	-0.25
Sq8	O=C1C(NCC2=CC(C(F)(F)F)=CC(C(F)(F)F)=C2)=C(N[C@@H]([C@@H]3[N@](CC4)C[C@@H](C=C)[C@H]4C3)C5=C(C=C(OC)C=C6)C6=NC=C5)C1=O	11.03	-0.24
Sq9	O=C1C(NC2=CC=C(C(F)(F)F)C=C2)=C(N[C@@H]([C@@H]3[N@](CC4)C[C@@H](C=C)[C@H]4C3)C5=C(C=CC=C6)C6=NC=C5)C1=O	12.18	-0.37
Sq10	O=C1C(NC2=CC(C(F)(F)F)=CC(C(F)(F)F)=C2)=C(N[C@@H]([C@@H]3[N@](CC4)C[C@@H](C=C)[C@H]4C3)C5=C(C=CC=C6)C6=NC=C5)C1=O	11.03	-0.43
Sq12	O=C1C(NC2=CC=C(C(F)(F)F)C=C2)=C(N[C@H]3[C@H](N(C)C)CCCC3)C1=O	13.1	0.09
Sq14	O=C1C(NC2=CC=C(C(F)(F)F)C=C2)=C(N[C@H]3[C@H](N4CCCCC4)CCCC3)C1=O	13.3	0.11
Sq15	O=C1C(NC2=CC(C(F)(F)F)=CC(C(F)(F)F)=C2)=C(N[C@H]3[C@H](N4CCCCC4)CCCC3)C1=O	11.87	-0.08
Sq17	O=C1C(C(N[C@@H]2CCCC[C@H]2NC3=C(NC4=CC(C(F)(F)F)=CC(C(F)(F)F)=C4)C(C3=O)=O)=C1NC5=CC(C(F)(F)F)=CC(C(F)(F)F)=C5)=O	10.78	-0.68
Sq18	O=C1C(NC2=CC(C(F)(F)F)=CC(C(F)(F)F)=C2)=C(N[C@@H](C3=CC=CC=C3)[C@@H](N(C)C)C4=CC=CC=C4)C1=O	11.42	-0.16
Sq19	O=C1C(NC2=CC(C(F)(F)F)=CC(C(F)(F)F)=C2)=C(N[C@@H](CC3=CC=CC=C3)CN(CC)CC)C1=O	11.57	-0.20
Sq20	O=C1C(NC2=CC=C(C(F)(F)F)C=C2)=C(NC3=CC=C(C(F)(F)F)C=C3)C1=O	9.8	-0.53
Sq21	O=C1C(NC2=CC(C(C)(C)C)=CC(C(C)(C)C)=C2)=C(NC3=CC(C(C)(C)C)=CC(C(C)(C)C)=C3)C1=O	12.1	0.22
Sq22	O=C1C(NC2=CC(C(F)(F)F)=CC(C(F)(F)F)=C2)=C(NCCCCCCCCCCCCC)C1=O	11.5	-0.12
ThSq1	S=C1C(NC2=CC=CC=C2)=C(NC3=CC=CC=C3)C1=S	7.3	-0.56
ThSq2	S=C1C(NC2=CC=C(C(F)(F)F)C=C2)=C(NC3=CC=C(C(F)(F)F)C=C3)C1=S	5.3	-1.05
ThSq3	S=C1C(NC2=CC(C(F)(F)F)=CC(C(F)(F)F)=C2)=C(NC3=CC(C(F)(F)F)=CC(C(F)(F)F)=C3)C1=S	4.9	-1.25
ThSq4	S=C1C(NC2=CC(C(C)(C)C)=CC(C(C)(C)C)=C2)=C(NC3=CC(C(C)(C)C)=CC(C(C)(C)C)=C3)C1=S	7.7	-0.49
ThSqO1	O=C1C(NC2=CC=C(C(F)(F)F)C=C2)=C(NC3=CC=C(C(F)(F)F)C=C3)C1=S	6.6	-1.04

3.2 Structures generation and analysis

The 1,593 DHBD catalysts in the seed and CSD-enriched databases were analyzed using substructure searches targeting 7 different hydrogen-bond-donor moieties [(thio)urea, (thio)squaramide, deltamide, croconamide, and sulfamide]. Their SMILES strings were used as input for RDKit;[7] covalent bonds starting from and not ending in an atom in the DHBD unit were cut to collect a subset of 694 unique substituents (after duplicate removal through canonicalization of the underlying SMILES strings). The 7 cores were then combined systematically with the 694 substituents to yield a combinatorially enriched database of 1,678,287 unique DHBDs. Using RDKit, 3D structures were generated, and 1,573,015 were successfully optimized at the GFN2-xTB level using xTB.[13] A subset of 7,000 structures was constructed by randomly selecting 1,000 structures per core with θ approximately < 90°.

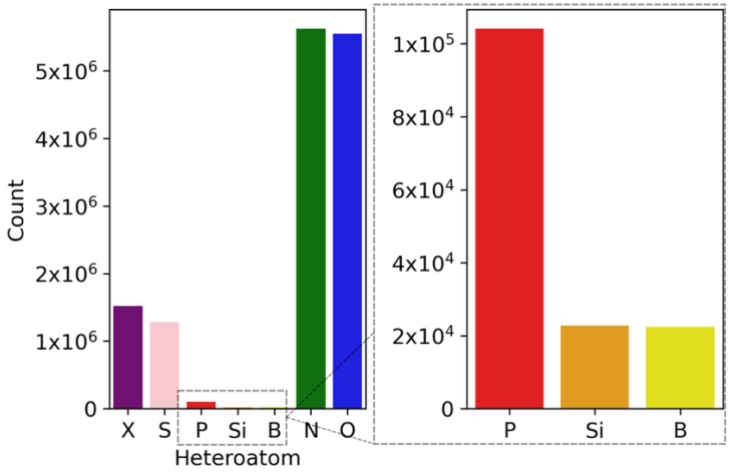

Figure S14. Distribution histograms of heteroatom types (X = halogen) in OSCAR!(DHBD) (1,573,015 structures).

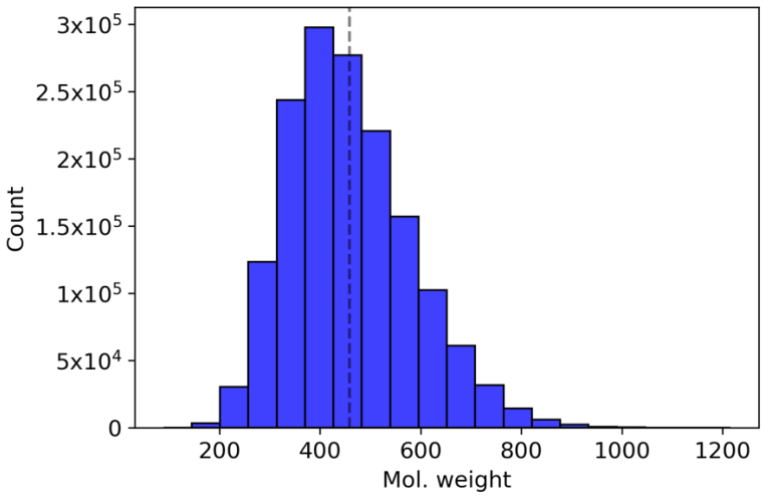

Figure S15. Distribution histograms of molecular weights in OSCAR!(DHBD) (1,573,015 structures).

Table S7. Count of functional groups (FG) present in OSCAR!(DHBD) (1,573,015 structures).

Atom type	FG	Description	OSCAR!(DHBD)
C	$R_3C–CR_3$	Alkane	6715519
C	$R_2C=CR_2$	Alkene	239983
C	RC#CR	Alkyne	50712
C	$R_2C=C=CR_2$	Allene	0
N	NR_3	Amine	2582908
N	C=N	Imine	1818215
N	C(=O)N	Amide	1463684
N	$R_2C=CNR_2$	Enamine	57361
N	NR_4^+	Ammonium	3938
N	$R_2C=NR_2^+$	Iminium	0
N	N=N	Azo	63649

N	N=N=N	Azide	4631
N	N–B	Amine-borane	0
N	RC#N	Nitrile	71520
N	N–N	Hydrazine	18441
O	C–O–C	Ether	572677
O	C(C=O)C	Ketone	1431358
O	RC(=O)OR	Carboxylic	457579
O	C=N(O)–R	Nitrone	214724
O	RNO_2	Nitro	100380
O	N–O	Other	50561
O	$ROCO_2R$	Carbonate	0
O	RO–OR	Peroxide	0
O	RO^-	Alkoxide	3381
O	$RCOO^-$	Carboxylate	0
O	RN=O	Nitroso	4650
O	N=C=O	Isocyanate	3090
S	CSN	Thioamide	478802
S	S–N	Thioamine	594368
S	SO_2R_2	Sulfone	360753
S	C–S–C	Thioether	226059
S	C–CS–C	Thiocarbonyl	415330
S	SO_3R	Sulfonate	9224
S	S–P	Thiophosphine	0
S	SOR_2	Sulfoxide	9163
S	CS_2R	Dithiocarboxylic	4638
S	COSR	Thiocarboxylic	0
S	SO_2R	Sulfinate	0
S	CSOR	Thiocarboxylic	0
P	PR_3	Phosphine	22493
P	PO_4	Phosphate	0
P	P=C	Ylide	26877
P	POR_3	Phosphine oxide	13924
P	PO_3N	Phosphoramidate	0
P	P–O	Phosphoxide	4234
P	P–N	Phosphinamine	11831
P	P=N	Phosphazene	8959
P	PO_3R	Phosphonate	55396
P	PO_2N	Phosphoramidite	0
X	C–X	Carbon halide	1495796
X	F	Fluorine	899978
X	Cl	Chlorine	523570
X	Br	Bromine	81494
X	S–X	Sulfur halide	0
X	B–X	Boron halide	9246
X	N–X	Nitrogen halide	0
B	BOR_2	Borninate	17811
B	BR_3	Borane	4623
B	BO_2R	Boronate	0
Ring	6-mem-ring		20724283
Ring	5-mem-ring		4765851
Ring	7-mem-ring		146901
Ring	4-mem-ring		1756620
Ring	3-mem-ring		775767

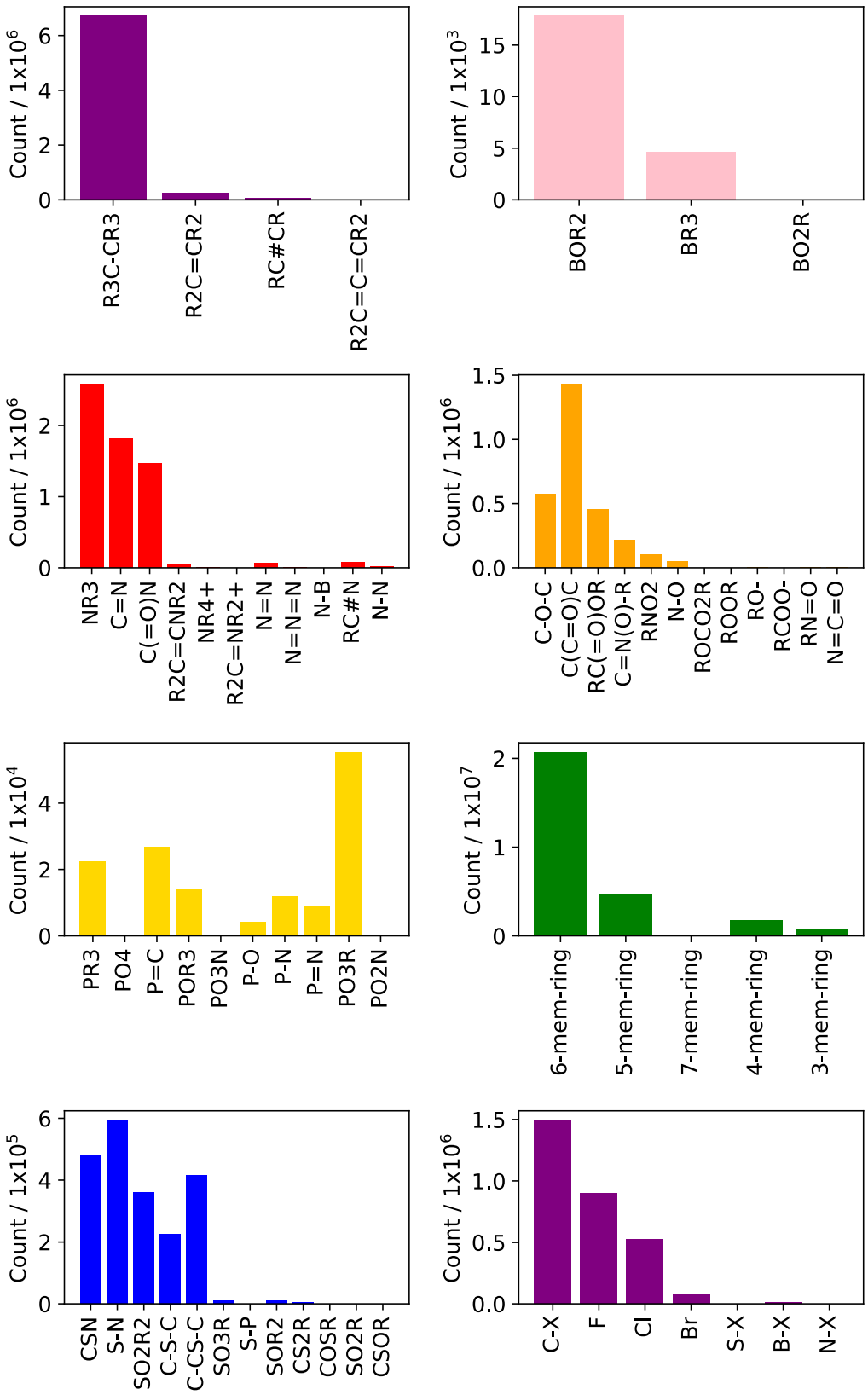

Figure S16. Bar plots of functional groups present in OSCAR!(DHBD) (data from Table S7).

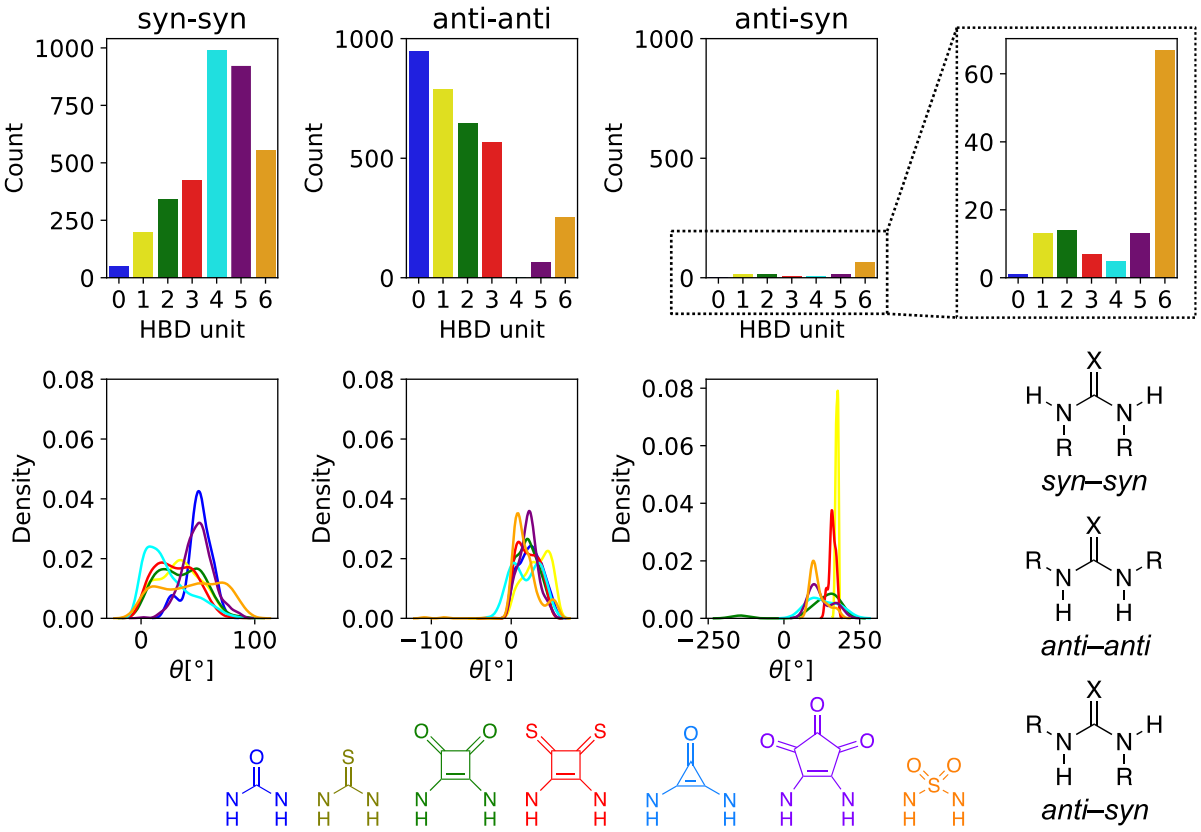

Figure S17. (Top) Distribution histograms of HBD units in the *syn-syn*, *anti-anti*, or *anti-syn* conformation (0 = ureas, 1 = thioureas, 2 = squaramides, 3 = thiosquaramides, 4 = deltamides, 5 = croconamides, 6 = sulfamides). (Bottom) Distribution plots of θ for each HBD unit in the *syn-syn*, *anti-anti*, or *anti-syn* conformation.

Table S8. Count of HBD units in the *syn-syn*, *anti-anti*, or *anti-syn* conformation.

Conformation	HBD unit	Count
syn–syn		**3482**
	Urea	50
	Thiourea	196
	Squaramide	341
	Thiosquaramide	426
	Delatamide	991
	Croconamide	922
	Sulfamide	556
anti–anti		**3267**
	Urea	949
	Thiourea	789
	Squaramide	645
	Thiosquaramide	566
	Delatamide	2
	Croconamide	64
	Sulfamide	252
anti–syn		**120**
	Urea	1
	Thiourea	13
	Squaramide	14
	Thiosquaramide	7
	Delatamide	5
	Croconamide	13
	Sulfamide	67

Figure S17 and Table S8 show the distribution of the 7 hydrogen-bonding units in OSCAR!(DHBD) according to their conformation (*anti-anti*, *syn-syn*, or *anti-syn*). Most of the ureas (dark blue, 95%) and thioureas (yellow, 80%) exist as *anti-anti* conformers. This is in agreement with the result by Paton *et al.*, who mined and analyzed diaryl(thio)ureas from CSD.[14] The number of HBD units in the *syn-syn* conformation increases in the order ureas < thioureas < squaramides < thiosquaramides < deltamides (light blue), while the opposite trend is found for the *anti-anti* configuration. Indeed, almost all of the deltamides (and over 92% of the croconamides) are in the *syn-syn* form. Only a small number of catalysts in OSCAR!(DHBD) exists as *anti-syn* conformers (as opposed to the CSD-extracted dataset, *cf* Figure 7D).

Analysis of the distribution of θ values (Figure S17, bottom) shows that, in the *syn-syn* configuration, the HNNH dihedral of most deltamides (light blue species) is closer to 0°. Ureas and croconamides (dark blue and purple curves) have a narrower distribution of θ's centered at higher values than (thio)squaramides and thioureas, while sulfamides (orange curve) are equally distributed over a wide range of θ's. In the *anti-anti* configuration, most DHBDs have a similar distribution of dihedral angles (max. ~20°), apart from thioureas who, on average, have slightly larger θ's (30°).

4. Conformational Analysis

In its current form, OSCAR contains one representative conformer of each organocatalyst from which stereoelectronic parameters are determined. That is because molecular geometries were optimized directly from the crystal structure extracted from CSD or from the Cartesian coordinates generated from the corresponding SMILES string using the OpenBabel implementation[15] of the Merck Molecular Force Field method (MMFF94).[16–20]

To assess the impact of accounting for the conformational space of each catalyst on the key molecular descriptors provided in OSCAR, conformational analysis was performed on three diverse entries from OSCAR!(NHC) and three from OSCAR!(DHBD). Structures were generated using CREST[21,22] and 10 geometries were selected based on RMSD clustering, followed by DFT optimization and single-point energy computations (at the ωB97X-D/Def2-TZVP//B97-D/Def2-TZVP level). Results are provided in Table S9 and Table S10. As can be evinced from Table S11 and Table S12, and in agreement with previous reports,[23,24] electronic properties are less sensitive to the conformational ensemble than steric ones. While θ is more sensitive to the flexibility of the DHBD organocatalysts, there is good agreement between the reported $\%V_{buried}$ value and the Boltzmann-weighted one. Overall, while OSCAR could be refined by providing physiochemical descriptors based on representative conformer ensembles,[24] the structures mined from CSD or generated from SMILES are a good basis for the reactivity indices provided, especially for the general exploration of wider regions of organocatalyst space.

Table S9. Selected conformers of structures in OSCAR!(NHC).

Structure	ΔE [kcal/mol]	*N*-index [1/eV]	$\%V_{buried}$
00000352_Conf11	0.1	0.95	68.8
00000352_Conf12	0.0	0.95	69.0
00000352_Conf16	1.3	0.91	52.3
00000352_Conf18	1.3	0.91	52.4
00000352_Conf19	1.5	0.91	40.3
00000352_Conf2	0.1	0.94	71.3
00000352_Conf4	1.5	0.91	42.8
00000352_Conf7	1.5	0.91	42.8
00000352_Conf8	1.3	0.91	52.4
00000352_Conf9	0.0	0.95	69.0
00002290_Conf11	2.0	3.20	30.2
00002290_Conf13	3.2	3.15	34.7
00002290_Conf14	2.7	3.17	36.0
00002290_Conf1	0.0	3.15	29.5
00002290_Conf2	0.0	3.15	30.4
00002290_Conf3	0.0	3.15	28.1
00002290_Conf5	0.0	3.15	29.9
00002290_Conf7	1.7	3.14	34.8
00002290_Conf8	1.0	3.14	29.6
00002290_Conf9	1.1	3.10	41.1
00006260_Conf11	3.8	1.37	31.5
00006260_Conf14	2.3	1.27	39.2
00006260_Conf16	2.8	1.26	36.4
00006260_Conf17	2.3	1.27	39.2
00006260_Conf19	2.8	1.26	36.5
00006260_Conf2	0.0	1.29	31.0
00006260_Conf4	0.0	1.29	31.9
00006260_Conf5	0.0	1.29	31.9
00006260_Conf6	3.8	1.37	31.4
00006260_Conf7	3.8	1.37	31.6

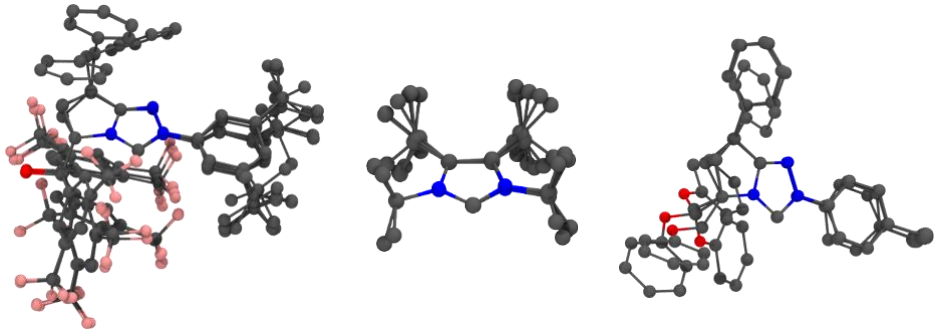

Figure S18. Ensembles of conformations of selected carbenes from OSCAR!(NHC).

Table S10. Selected conformers of structures in OSCAR!(DHBD).

Structure	**ΔE [kcal/mol]**	**ε_{LUMO} [eV]**	***θ* [°]**
00243771_Conf10	1.1	2.037	10.2
00243771_Conf15	1.9	2.074	12.0
00243771_Conf1	2.0	1.990	15.9
00243771_Conf20	2.1	2.063	1.6
00243771_Conf30	6.0	2.425	38.7
00243771_Conf33	6.0	2.425	38.4
00243771_Conf3	0.0	2.024	13.0
00243771_Conf40	6.1	2.263	1.1
00243771_Conf44	7.3	2.339	0.3
00243771_Conf6	2.0	2.124	16.4
00499240_Conf118	24.8	–0.119	106.8
00499240_Conf12	0.0	0.289	171.8
00499240_Conf145	30.7	0.452	88.7
00499240_Conf16	0.5	0.292	173.7
00499240_Conf19	0.2	0.185	164.2
00499240_Conf1	0.3	0.319	175.5
00499240_Conf42	9.9	0.133	162.2
00499240_Conf4	0.5	0.292	173.7
00499240_Conf61	9.9	0.133	162.2
00499240_Conf99	18.8	–0.016	173.9
01592890_Conf12	6.8	–1.450	18.7
01592890_Conf13	1.0	–1.293	10.6
01592890_Conf19	0.0	–1.262	22.4
01592890_Conf20	1.9	–1.257	3.0
01592890_Conf23	2.1	–1.192	147.7
01592890_Conf25	9.9	–1.424	33.7
01592890_Conf3	0.6	–1.188	157.2
01592890_Conf45	7.7	–0.908	161.3
01592890_Conf49	7.6	–1.102	139.8
01592890_Conf53	6.7	–1.060	158.8

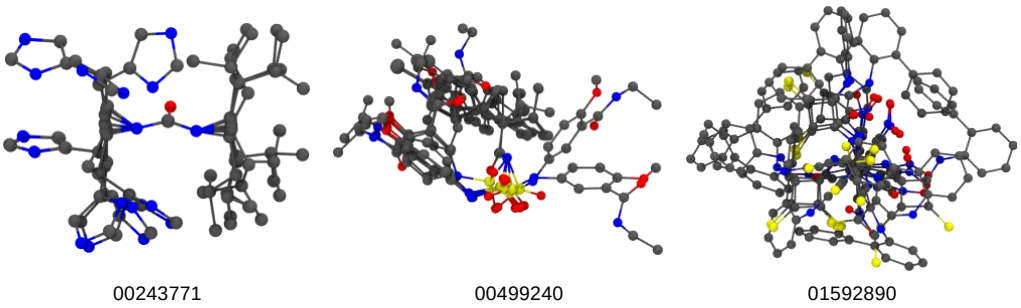

Figure S19. Ensembles of conformations of selected carbenes from OSCAR!(DHBD).

Table S11. Stereoelectronic descriptors of selected structures from OSCAR!(NHC).

Structure	*N*-index [1/eV]				%V_{buried}			
	Reported[a]	w[b]	min	max	Reported[a]	w[b]	min	max
00000352	0.88	0.94	0.91	0.95	72.0	66.6	40.3	71.3
00002290	3.11	3.15	3.10	3.20	28.3	30.0	28.1	41.1
00006260	1.27	1.29	1.26	1.37	33.7	31.7	31.0	39.2

[a]Value used to construct the chemical space map in Figure 6A. [b]Boltzmann-weighted value (298 K).

Table S12. Stereoelectronic descriptors of selected structures from OSCAR!(DHBD).

Structure	ε_{LUMO} [eV]				θ [°]			
	Reported[a]	w[b]	min	max	Reported[a]	w[b]	min	max
00243771	2.44	2.03	1.99	2.43	3.0	12.5	0.3	38.7
00499240	0.00	0.27	–0.12	0.45	97.0	171.3	88.7	175.5
01592890	–1.29	–1.25	–1.45	–0.91	10.7	52.4	3.0	161.3

[a]Value used to construct the chemical space map in Figure 6B. [b]Boltzmann-weighted value (298 K).

5. Structures and Descriptors Availability

5.1 Additional reactivity indices

The reactivity indices E_{index}, N_{index}, and N_{rel} are discussed in the main paper. Additional descriptors, based on the work by Kang *et al.*,[25] are also provided with the data on the Materials Cloud (*vide infra*). These "global reactivity indices" for electrophilicity (*E*-GRI) and nucleophilicity (*N*-GRI) are calculated according to equations (S1) and (S2):

$$E\text{-GRI} = \log_{10}\left(\sum_m 10^{\text{EI}_\text{r}}\right) \quad \text{(S1)}$$

$$N\text{-GRI} = \log_{10}\left(\sum_m 10^{\text{NI}_\text{r}}\right) \quad \text{(S2)}$$

$$\text{EI}_\text{r} = |\rho_\text{s}^{n+1}(r)| \times E\text{-index} \quad \text{(S3)}$$

$$\text{NI}_\text{r} = |\rho_\text{s}^{n-1}(r)| \times N_\text{rel} \quad \text{(S4)}$$

where m is the number of atoms in the catalyst, and EI_r and NI_r are the local indices for electrophilicity and nucleophilicity at each atom r, expressed as the product of E-index and N_{rel} and the atomic spin density $[\rho_\text{s}(r)]$ of the n+1 or n–1-electron species.

Inspired by Mayr's linear free-energy relationship $\log_{10}(k_{\text{reaction}})_{20\,^\circ\text{C}} = s_\text{N}(E + N)$,[26–28] theoretical reaction rate constants for electrophilic and nucleophilic attack were estimated by substituting the experimental electrophilicity E and nucleophilicity N parameters with E-GRI and N-GRI, as defined above (assuming $s_\text{N} \approx 1$).[25]
Thus, the rate for an organocatalyst undergoing nucleophilic attack [*i.e.*, acting as an electrophile, $\log_{10}(k_\text{E})$], or electrophilic attack [*i.e.*, acting as a nucleophile, $\log_{10}(k_\text{N})$] are defined as follows:

$$\log_{10}(k_\text{E}) \propto E\text{-GRI(organocatalyst)} + N\text{-GRI}(\text{OH}^-) \quad \text{(S5)}$$

$$\log_{10}(k_\text{N}) \propto E\text{-GRI}(\text{CH}_3^+) + N\text{-GRI(organocatalyst)} \quad \text{(S6)}$$

where the methyl cation and hydroxyl anion have been chosen as model electrophile and nucleophile, respectively.

5.2 Open shell computations and alternative conceptual DFT descriptors

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 4,000 catalysts in the seed and CSD-extracted dataset and for the 8,622 carbenes in OSCAR!(NHC).
These energies provide an alternative way of estimating the organocatalysts' vertical ionization potential (IP) and electron affinity (EA) according to equations (S7) and (S8):

$$\mathrm{IP} = \mathrm{E}(n\text{-}1) - \mathrm{E}(n) \tag{S7}$$

$$\mathrm{EA} = \mathrm{E}(n) - \mathrm{E}(n\text{+}1) \tag{S8}$$

Consequently, the conceptual DFT descriptors hardness (η), electronegativity (χ), electrophilicity (ω, or *E*-index) and nucleophilicity (*N*-index) can be calculated as follows:[29]

$$\eta = \frac{(\mathrm{IP} - \mathrm{EA})}{2} \tag{S9}$$

$$\chi = \frac{(\mathrm{IP} + \mathrm{EA})}{2} \tag{S10}$$

$$\omega = \frac{\chi^2}{2\eta} \tag{S11}$$

$$N_{\mathrm{index}} = \frac{1}{\omega} \tag{S12}$$

These energy-based electrophilicity and nucleophilicity descriptors (*i.e.*, calculated from the single-point energy computations) constitute an alternative to the frontier molecular orbital energy-based descriptors presented in the main paper.

5.3 Data on the Materials Cloud

XYZ structures and molecular descriptors can be found at https://doi.org/10.24435/materialscloud:gy-3h and can be visualized interactively with Chemiscope.[30] The data is separated into 5 sets, each corresponding to a csv file and a set of XYZ coordinates:

1. Seed_CSD: 4,000 DFT-optimized structures from the seed and CSD-extracted datasets and corresponding molecular descriptors. The "Type-code" in the csv file corresponds to the catalytic motif, according to Table S13;
2. DLPNO-CCSD: subset of 2,060 structures from the seed and CSD-extracted sets with parameters computed at the EOM-DLPNO-CCSD/cc-pVTZ level;
3. OSCAR_NHC: 8,622 DFT-optimized structures from OSCAR!(NHC) and corresponding molecular descriptors;
4. OSCAR_DHBD_xTB: 1,573,015 xTB-optimized structures from OSCAR!(DHBD). The corresponding SMILES strings are listed in SMILES_OSCAR_DHBD_xTB.csv;
5. OSCAR_DHBD_DFT: subset of 6,994 DFT-optimized structures from OSCAR!(DHBD) and corresponding molecular descriptors.

Table S13. List of catalytic motifs and corresponding numbering.

Type-code	Catalytic motif
1	SHBD
2	DHBD
3	PA
4	BB
5	LA
6	LB
7	NHC
8	Aminocat.

6. References

1 S. Vela, R. Laplaza, Y. Cho and C. Corminboeuf, cell2mol: encoding chemistry to interpret crystallographic data, *npj Comput. Mater.*, 2022, **8**, 188.
2 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.
3 A. S. Christensen, F. A. Faber, B. Huang, L. A. Bratholm, A. Tkatchenko, K.-R. Müller and O. A. von Lilienfeld, 2017, QML: A Python Toolkit for Quantum Machine Learning, https://github.com/qmlcode/qml.
4 L. van der Maaten and G. Hinton, Visualizing Data using t-SNE, *J. Mach. Learn. Res.*, 2008, **9**, 2579–2605.
5 L. McInnes, J. Healy and J. Melville, UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, *arXiv:1802.03426v3*.
6 D. Rogers and M. Hahn, Extended-Connectivity Fingerprints, *J. Chem. Inf. Model.*, 2010, **50**, 742–754.
7 RDKit: open-source chemoinformatics and machine learning. http://www.rdkit.org.
8 N. Halko, P.-G. Martinsson and J. A. Tropp, Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions, *arXiv:0909.4061*.
9 A. Fabrizio, B. Meyer and C. Corminboeuf, Machine learning models of the energy curvature vs particle number for optimal tuning of long-range corrected functionals, *J. Chem. Phys.*, 2020, **152**, 154103.
10 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.
11 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.
12 Q. Yang, Y. Li, J.-D. Yang, Y. Liu, L. Zhang, S. Luo and J.-P. Cheng, Holistic Prediction of the pKa in Diverse Solvents Based on a Machine-Learning Approach, *Angew. Chem. Int. Ed.*, 2020, **59**, 19282–19291.
13 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.
14 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.
15 N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch and G. R. Hutchison, Open Babel: An open chemical toolbox, *J. Cheminform.*, 2011, **3**, 33.
16 T. A. Halgren, Merck molecular force field .1. Basis, form, scope, parameterization, and performance of MMFF94, *J. Comput. Chem.*, 1996, **17**, 490–519.
17 T. A. Halgren, Merck molecular force field .2. MMFF94 van der Waals and electrostatic parameters for intermolecular interactions, *J. Comput. Chem.*, 1996, **17**, 520–552.
18 T. A. Halgren, Merck molecular force field .3. Molecular geometries and vibrational frequencies for MMFF94, *J. Comput. Chem.*, 1996, **17**, 553–586.
19 T. A. Halgren and R. B. Nachbar, Merck molecular force field .4. Conformational energies and geometries for MMFF94, *J. Comput. Chem.*, 1996, **17**, 587–615.
20 T. A. Halgren, Merck molecular force field .5. Extension of MMFF94 using experimental data, additional computational data, and empirical rules, *J. Comput. Chem.*, 1996, **17**, 616–641.
21 S. Grimme, F. Bohle, A. Hansen, P. Pracht, S. Spicher and M. Stahn, Efficient Quantum Chemical Calculation of Structure Ensembles and Free Energies for Nonrigid Molecules, *J. Phys. Chem. A*, 2021, **125**, 4039–4054.
22 P. Pracht, F. Bohle and S. Grimme, Automated exploration of the low-energy chemical space with fast quantum chemical methods, *Phys. Chem. Chem. Phys.*, 2020, **22**, 7169–7192.
23 L. Wilbraham, E. Berardo, L. Turcani, K. E. Jelfs and M. A. Zwijnenburg, High-Throughput Screening Approach for the Optoelectronic Properties of Conjugated Polymers, *J. Chem. Inf. Model.*, 2018, **58**, 2450–2459.
24 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.
25 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.
26 H. Mayr and M. Patz, Scales of Nucleophilicity and Electrophilicity: A System for Ordering Polar Organic and Organometallic Reactions, *Angew. Chem. Int. Ed.*, 1994, **33**, 938–957.
27 H. Mayr, T. Bug, M. F. Gotta, N. Hering, B. Irrgang, B. Janker, B. Kempf, R. Loos, A. R. Ofial, G. Remennikov and H. Schimmel, Reference Scales for the Characterization of Cationic Electrophiles and Neutral Nucleophiles, *J. Am. Chem. Soc.*, 2001, **123**, 9500–9512.
28 H. Mayr and A. R. Ofial, Do general nucleophilicity scales exist?, *J. Phys. Org. Chem.*, 2008, **21**, 584–595.

29 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.
30 G. Fraux, R. K. Cersonsky and M. Ceriotti, Chemiscope: interactive structure-property explorer for materials and molecules, *J. Open Source Softw.*, 2020, **5**, 2117.