Supplementary information

Contents

S1 Lewis acidity metrics benchmark

As explained in the manuscript, we employed a computed quantity to account for the Lewis acidity of boron derivatives, the fluoride ion affinity (FIA). However, this choice results from a preliminary study on the different possible scales to account for Lewis acidity. This study is detailed here.

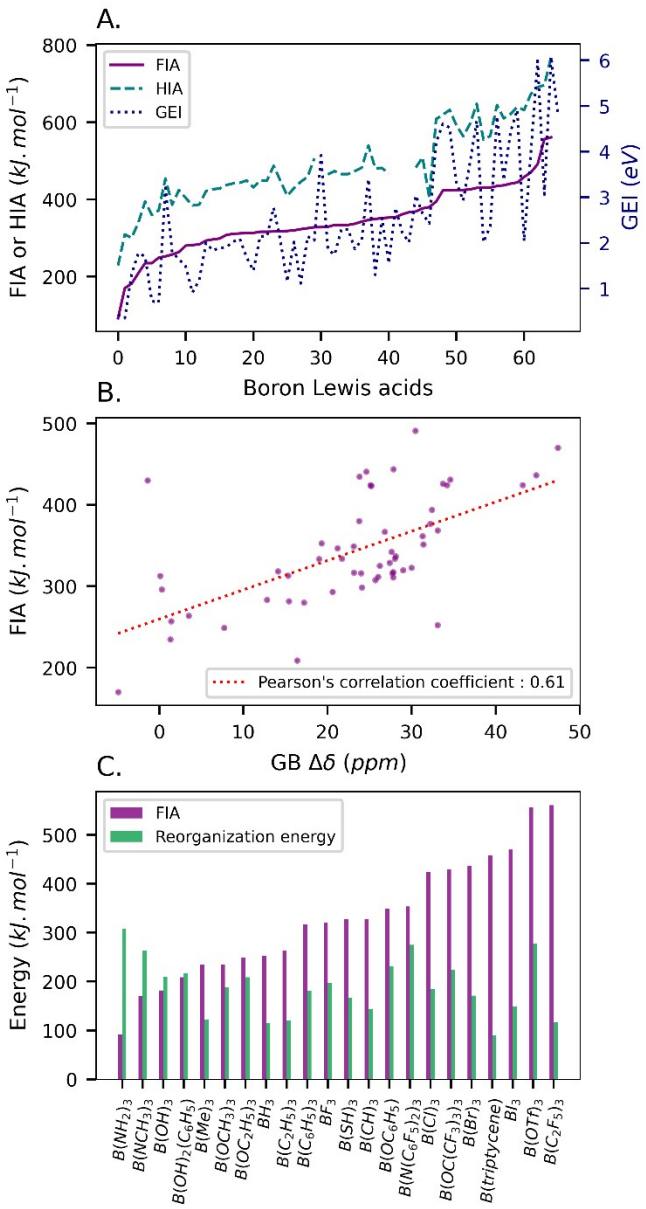


Figure S1 : Lewis acidity description: A. correlation between FIA, HIA and GEI, B. correlation between FIA and GB Δδ(^{31}P), C. comparison of FIA and the corresponding reorganization energy for various common Lewis acids.

Traditional Lewis acidity scales typically involve measuring the interaction of a Lewis acid with a specific base for comparison. Among these scales, FIA is particularly well-established, as the fluoride ion offers significant advantages as a Lewis base. Its small size and low polarizability minimize steric hindrance and mitigate interfering effects such as charge transfer, π-back-bonding, and dispersion forces.[1] Experimental determination of FIA requires precise conditions[2] ; hence, it is frequently calculated computationally using density functional theory (DFT), ensuring consistency.

According to the Hard and Soft Acids and Bases (HSAB) theory,[3] the fluoride ion is classified as a hard base. Consequently, we also considered the hydride ion affinity (HIA) as an alternative Lewis acidity scale, given that the hydride ion is a soft Lewis base, although it is less commonly employed.

Both FIA and HIA are thermodynamic measures of Lewis acidity, based on the standard enthalpy change of Lewis adduct formation. In addition to these thermodynamic Lewis acidity scales,[4] it is also possible to assess a compound's Lewis acidity through analysis of its electronic structure, known as intrinsic Lewis acidity scales,[4] which require only minimal quantum calculation. One example is the global electrophilicity index (GEI),[5] that can easily be calculated from the energies of the molecule's frontier orbitals :

$$GEI = \frac{\chi^2}{2\eta}$$

where χ represents the molecule's global electronegativity and η denotes its hardness. The GEI measures a molecule's ability to accept electrons and has the advantage of being independent of the Lewis base. However, it is linked to reaction kinetics, contrasting with the conventional thermodynamic definition of Lewis acidity. Refer to part *S2 Computational methods* for details.

These theoretical scales should correlate with the experimental ones, even though experimental data can be noisy. Typical experimental methodologies involve measuring changes in the physicochemical properties of a probe base molecule upon binding with the studied Lewis acid. Two preferred bases are triethylphosphine oxide -or another phosphine oxide-, using the Gutmann-Beckett (GB) method (Figure S2.A),[4,6] and crotonaldehyde,[7] following Child's method (Figure S2.B). For both methods, the approach involves comparing nuclear magnetic resonance (NMR) chemical shifts, either of the terminal proton in crotonaldehyde or of the phosphorus atom in triethylphosphine oxide, when free in solution and when bound to the Lewis acid. The difference in NMR shifts, denoted as Δδ, serves as a consistent metric for comparing different Lewis acids.[4]

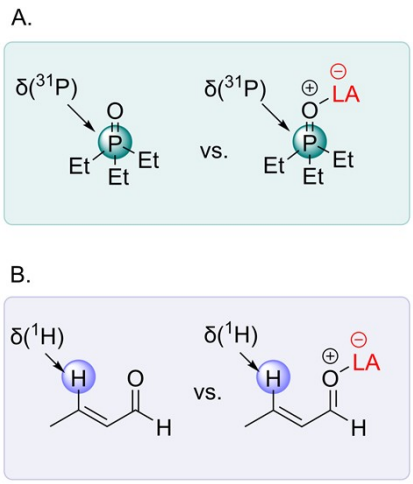


Figure S2 : NMR spectroscopy methodologies to determine the Lewis acidity of a Lewis acid (LA): A. Gutmann-Beckett method. B. Childs method.

Comparing Lewis acidity scales is complex, particularly as experimental measurements of Lewis acidity are prone to numerous artifacts. Challenges such as poor solubility or incomplete adduct formation can interfere with the accurate

measurement of phosphorus chemical shift.[4] Moreover, considering the need for a large dataset of boron Lewis acids to build a machine learning (ML) database, having a computable quantity that can be consistently accessed across molecules is advantageous. To ensure alignment with the experimental scale, we first calculated the fluoride ion affinity (FIA), hydride ion affinity (HIA), and global electrophilicity index (GEI) for various Lewis acids (56 molecules) where the GB $\Delta\delta(^{31}P)$ was reported in literature (refer to Figure S1.B). The computed values and NMR data collected from literature are presented in Table S1 and analyzed in a pair-plot (Figure S3).

FIA or HIA and GEI are somewhat correlated but not perfectly (Figure S1.A); Pearson's correlation coefficient (Pearson's r) between FIA and GEI is 0.76, indicating that electrophilicity and Lewis acidity follow similar trends but are distinct concepts. FIA and HIA exhibit a strong correlation (Pearson's r = 0.95), allowing to use them interchangeably for comparing boron Lewis acids. Since the FIA is a global Lewis acidity scale and shows the best correlation (Pearson's r = 0.61) with the Gutmann-Beckett $\Delta\delta(^{31}P)$ among evaluated metrics (Figure S1.B, Figure S3 and Table S2), we selected it as the metric for labeling the Lewis acids of our database. The discrepancies between the FIA and GB scales primarily stem from steric effects, which are not fully captured by FIA. For instance, tris(pentachlorophenyl)borane ($B(C_6Cl_5)_3$) is predicted to be a strong Lewis acid with a computed FIA of 429.7 kJ.mol-1 ; however, its effective Lewis acidity measured by the GB method is only -1.4 ppm, attributed to steric hindrance.[8]

Berionni et al. have demonstrated that altering the typical planar trigonal boron coordination environment can significantly enhance the Lewis acidity of boranetriptycenes.[9] We have thus calculated the reorganization energy of Lewis acids, which quantifies the energy difference between the Lewis acid's geometry upon adduct formation (removing the fluoride ion) and its initial equilibrium geometry (Figure S1.C). As expected, Berionni's borane triptycene[9] achieves the smaller reorganization energy, due to the preorganized geometry of the ligand.

Although weak, the negative correlation between reorganization energy and FIA suggests that lower reorganization energy may be associated with higher FIA (Pearson's r = -0.27, Figure S1.C). While reorganization energy significantly influences the Lewis acidity of a compound,[9] it is heavily dependent on the molecular structure. For a given ligand scaffold with various substitutions, the values fluctuate only slightly around the mean (Figure S4). Since our interest lies primarily in varying the substituents rather than the carbon backbone of the ligand, reorganization energy alone was not a suitable metric for our purposes. However, FIA, being the opposite of the reaction enthalpy of adduct formation, also incorporates the geometric reorganization energy associated with complexation, making it a more comprehensive and relevant parameter for assessing Lewis acidity.

SMILES	$\delta(^{31}P)$ (solvent)	$\Delta\delta(^{31}P)$	average $\Delta\delta(^{31}P)$	FIA (kJ.mol-1)	HIA (kJ.mol-1)	Reorganization energy (kJ.mol-1)	GEI (eV)
CCOB(OCC)OCC	48.7 (neat)[6]	7.7[6]	7.7	248.69	371.27	208.95	0.69
COB(OC)OC	48.1 (C6D6)[10]	1.3[10]	1.3	234.63	359.50	188.53	0.73
ClCCOB(OCCCl)OCCCl	55.1 (neat)[11]	14.1[11]	14.1	318.12	409.35	214.23	1.15
ClCCCOB(OCCCCl)OCCCCl	56.3 (neat)[11]	15.3[11]	15.3	312.98	431.19	221.68	1.39
ClCCCCOB(OCCCCCl)OCCCCCl	53.8 (neat)[11]	12.8[11]	12.8	283.23	386.44	237.24	1.19
ClC(Cl)COB(OCC(Cl)Cl)OCC(Cl)Cl	67.8 (neat)[11]	26.8[11]	26.8	366.71		223.02	2.01
ClC(Cl)(Cl)COB(OCC(Cl)(Cl)Cl)OCC(Cl)(Cl)Cl	73.2 (neat)[11]	32.2[11]	32.2	376.53	508.65	247.59	2.64

ClCC(CCl)OB(OC(CCl)CCl)OC(CCl)CCl	60.3 (neat)[11]	19.3[11]	19.3	352.71	466.71	217.16	1.55
BrCCOB(OCCBr)OCCBr	58.2 (neat)[11]	17.2[11]	17.2	279.96	403.56	248.73	1.49
ICCOB(OCCI)OCCI	61.6 (neat)[11]	20.6[11]	20.6	293.11	421.32		2.00
FC(F)(F)COB(OCC(F)(F)F)OCC(F)(F)F	71 (neat)[11]	30[11]	30	322.69	445.66		1.11
CC[Si](CC)(CC)OB(O[Si](CC)(CC)CC)O[Si](CC)(CC)CC	56.4 (neat)[12]	15.4[12]	15.4	281.58	384.07	235.02	0.91
Fc1c(F)c(F)c(B(c2c(F)c(F)c(F)c(F)c2F)c2c(F)c(F)c(F)c(F)c2F)c(F)c1F	77 (CD2Cl2)[13], 75.4 (C6D6)[13]	26.3[13], 29.4[13]	27.85	443.82	639.34		4.91
c1ccc(B(c2ccccc2)c2ccccc2)cc1	72.5 (C6D6)[10]	20.6[4], 25.7[10]	23.15	316.68	486.91	181.31	2.77
Fc1cc(F)c(F)c(B(c2c(F)c(F)cc(F)c2F)c2c(F)c(F)cc(F)c2F)c1F	77.3 (CD2Cl2)[14]	25.2[14]	25.2	422.97	619.13		4.61
Fc1cc(B(c2cc(F)c(F)c(F)c2F)c2cc(F)c(F)c(F)c2F)c(F)c(F)c1F	76.7 (CD2Cl2)[15]	24.6[15]	24.6	440.66	620.91		4.44
Fc1cccc(F)c1B(c1c(F)cccc1F)c1c(F)cccc1F	72.6 (CD2Cl2)[15]	21.2[15]	21.2	346.60	539.54		3.40
FC(F)(F)c1cc(B(c2cc(C(F)(F)F)cc(C(F)(F)F)c2)c2cc(C(F)(F)F)cc(C(F)(F)F)c2)cc(C(F)(F)F)c1	78.9(CD2Cl2)[16]	28.2[16]	28.2				4.82
Fc1c(F)c(F)c(B(c2c(F)c(F)c(F)c(F)c2F)c2c(Cl)c(Cl)c(Cl)c(Cl)c2Cl)c(F)c1F	75.8 (CD2Cl2)[8]	25.1[8]	25.1	424.08	631.72		4.53
Fc1c(F)c(F)c(B(c2c(Cl)c(Cl)c(Cl)c(Cl)c2Cl)c2c(Cl)c(Cl)c(Cl)c(Cl)c2Cl)c(F)c1F	74.5 (CD2Cl2)[8]	23.8[8]	23.8	434.46	644.34		4.77
FC(F)(F)c1cc(C(F)(F)F)c(Bc2c(C(F)(F)F)cc(C(F)(F)F)cc2C(F)(F)F)c(C(F)(F)F)c1	78.7(C6D6)[17]	32.4[17]	32.4	393.92	607.95		4.20
c1ccc(OB(Oc2ccccc2)Oc2ccccc2)cc1	69.45(C6D6)[18]	23.1[18]	23.1	349.03	480.62	231.64	1.30
Fc1c(F)c(F)c(OB(c2c(F)c(F)c(F)c(F)c2F)c2c(F)c(F)c(F)c(F)c2F)c(F)c1F	80(C6D6)[18]	33.7[18]	33.7	425.67	595.22		3.75
Fc1c(F)c(F)c(OB(Oc2c(F)c(F)c(F)c(F)c2F)c2c(F)c(F)c(F)c(F)c2F)c(F)c1F	80.5(C6D6)[18]	34.2[18]	34.2	424.26	563.69		2.93
Fc1c(F)c(F)c(OB(Oc2c(F)c(F)c(F)c(F)c2F)Oc2c(F)c(F)c(F)c(F)c2F)c(F)c1F	80.9(C6D6)[18]	34.6[18]	34.6	430.73	564.14		2.37
FB(F)F	80.9 (neat)[4]	29[4]	29	319.65	428.92	197.18	2.02
FB(F)OOOSC(F)(F)F	84.6 (CDCl3)[19]	33.1[19]	33.1	368.45	482.43		3.04
B		33.1[4]	33.1	252.30	453.88	114.25	3.21
CCB(CC)CC	51.9 (C6D6)[10]	1.9[4], 5.1[10]	3.5	263.54	424.45	120.61	1.64
ClB(Cl)Cl	88.7 (neat)[6]	47.7[6], 38.7[4]	43.2	424.23	591.88	184.33	3.33
BrB(Br)Br	90.3 (neat)[6]	49.3[6], 40.3[4]	44.8	436.41	610.00	170.92	3.38
IB(I)I	92.9 (neat)[6]	51.9[6], 42.9[4]	47.4	469.81	672.95	149.09	3.97
OB(O)c1ccccc1	63.2(C6D6)[10]	16.4[10]	16.4	208.73	339.53	216.78	1.78
c1ccc(B2Oc3ccccc3O2)cc1	70.5(C6D6)[20]	24.1[20]	24.1	298.45	427.99	188.20	1.89
Fc1ccccc1B1Oc2ccccc2O1	74.2(C6D6)[20]	27.8[20]	27.8	310.75	441.25		2.03
Fc1cccc(B2Oc3ccccc3O2)c1	72.4(C6D6)[20]	26[20]	26	311.66	442.95		2.09
Fc1ccc(B2Oc3ccccc3O2)cc1	72.1(C6D6)[20]	25.7[20]	25.7	307.55	437.03		1.90

Fc1ccc(B2Oc3ccccc3O2)c(F)c1	74.2(C6D6)[20]	27.8[20]	27.8	317.25	447.72		2.04
Fc1cccc(F)c1B1Oc2ccccc2O1	74.1 (C6D6)[20]	27.7[20]	27.7	316.04	448.68		2.15
Fc1cc(B2Oc3ccccc3O2)cc(F)c1F	74.4 (C6D6)[20]	28[20]	28	333.67	465.44		2.30
Fc1cc(F)c(B2Oc3ccccc3O2)c(F)c1	72.6 (C6D6)[20]	26.2[20]	26.2	325.11	457.94		2.14
Fc1c(F)c(F)c(B2Oc3ccccc3O2)c(F)c1F	77.8 (C6D6)[20]	31.4[20]	31.4	351.40	481.32		2.54
Fc1c(F)c(F)c(B2Cc3ccc4ccccc4c3-c3c (ccc4ccccc34)C2)c(F)c1F	72.7 (CD2Cl2)[13], 71.2 (C6D6)[13]	22[13], 25.5[13]	23.75	380.03	403.61		2.43
CC1(C)OB(c2ccccc2F)OC1(C)C	47.8 (C6D6)[20]	1.4[20]	1.4	257.02	385.94		1.79
Fc1c(F)c(C(F)(F)F)c(F)c(F)c1B(c1c(F)c(F) c(C(F)(F)F)c(F)c1F)c1c(F)c(F)c(C(F)(F)F)c(F)c1F	79.5 (CD2Cl2)[21], 77.7 (C6D6)[21]	29[21], 31.9[21]	30.45	490.99	690.76		5.99
Clc1c(Cl)c(Cl)c(B(c2c(Cl)c(Cl)c(Cl)c(Cl) c2Cl)c2c(Cl)c(Cl)c(Cl)c(Cl)c2Cl)c(Cl)c1Cl	50.7 (CD2Cl2)[8]	-1.4[8]	-1.4	429.70	647.31	157.18	4.67
CN(C)B(N(C)C)N(C)C		-4.9[4]	-4.9	169.83	309.15	262.83	0.35
Fc1cc(B2Oc3ccccc3O2)cc(F)c1F	67.9(C6D6)[22]	21.7[22]	21.7	333.67	465.44		2.30
Fc1c(F)c(F)c(B2Oc3cccc4cccc(c34)O2)c(F)c1F	77.5(C6D6)[22]	31.3[22]	31.3	361.31	495.77		2.22
Fc1cc(B2Oc3cccc4cccc(c34)O2)cc(F)c1F	73.8(C6D6)[22]	27.6[22]	27.6	342.18	481.96		2.06
Fc1cc(F)c(B2Oc3cccc4cccc(c34)O2)c(F)c1	74.3 (C6D6)[22]	28.1[22]	28.1	336.77	473.19		1.86
Fc1cccc(F)c1B1Oc2cccc3cccc(c23)O1	73.6 (C6D6)[22]	27.4[22]	27.4	328.77	464.73		1.85
Clc1cccc(Cl)c1B1Oc2cccc3cccc(c23)O1	65.2 (C6D6)[22]	19[22]	19	333.59	473.11	187.88	1.74
Cc1cc(C)c(B2Oc3cccc4cccc(c34)O2)c(C)c1	46.3 (C6D6) [22]	0.1[22]	0.1	312.71	448.76	176.74	1.68
Fc1cccc(F)c1B1Oc2ccccc2O1	70.2 (C6D6)[22]	24[22]	24	316.04	448.68		2.15
Cc1cc(C)c(B2Oc3ccccc3O2)c(C)c1	46.5 (C6D6)[22]	0.3[22]	0.3	296.11	426.28	163.71	1.83
C#CB(C#C)C#C				327.86		144.37	3.92
c1ccc2c(c1)B1c3ccccc3C2c2ccccc21				457.77	631.25	89.73	2.04
CB(C)C				234.60	394.51	122.14	1.71
FC(F)(F)C(F)(F)B(C(F)(F)C(F)(F)F)C(F)(F)C(F)(F)F				560.68	769.73	117.24	6.08
FC(F)(F)C(OB(OC(C(F)(F)F)(C(F)(F)F)C(F)(F)F)O C(C(F)(F)F)(C(F)(F)F)C(F)(F)F)(C(F)(F)F)C(F)(F)F				429.88	554.52	224.35	2.01
Fc1c(F)c(F)c(N(B(N(c2c(F)c(F)c(F)c(F)c2F)c2 c(F)c(F)c(F)c(F)c2F)N(c2c(F)c(F)c(F)c(F)c2F) c2c(F)c(F)c(F)c(F)c2F)c2c(F)c(F)c(F)c(F)c2F)c(F)c1F				353.84		275.20	2.77
NB(N)N				92.01	228.71	307.55	0.43
O=S(=O)(OB(OS(=O)(=O)C(F)(F)F)OS(=O)(=O)C(F) (F)F)C(F)(F)F				555.93	695.88	276.97	3.06
OB(O)O				181.47	300.20	209.07	1.31
SB(S)S				327.73	506.10	166.85	2.05

Table S1 : Lewis acidity data for a bunch of Lewis acids. GB Δδ(^{31}P) calculated from NMR spectroscopy data reported in the literature, computed FIA, reorganization energy and GEI at the DFT M06-2X/6-31G(d) level of theory. As δ(^{31}P) of the triethylphosphine oxide complexed with the Lewis acid were often reported in diverse solvents (mainly CD_2Cl_2 and C_6D_6) and sometimes in neat conditions, we calculated the Δδ(^{31}P) subtracting the value of δ(^{31}P) for the triethylphosphine oxide in the appropriate conditions (as described in the GB methodology)[6]: 41 ppm (neat), 50.7 ppm (CD_2Cl_2), 46.3 ppm (C_6D_6) or 51.5 ppm ($CDCl_3$). When Δδ(^{31}P) was also reported, this value was used instead. Then, the average value of all the Δδ(^{31}P) reported for a same molecule was calculated and used for the following correlations.

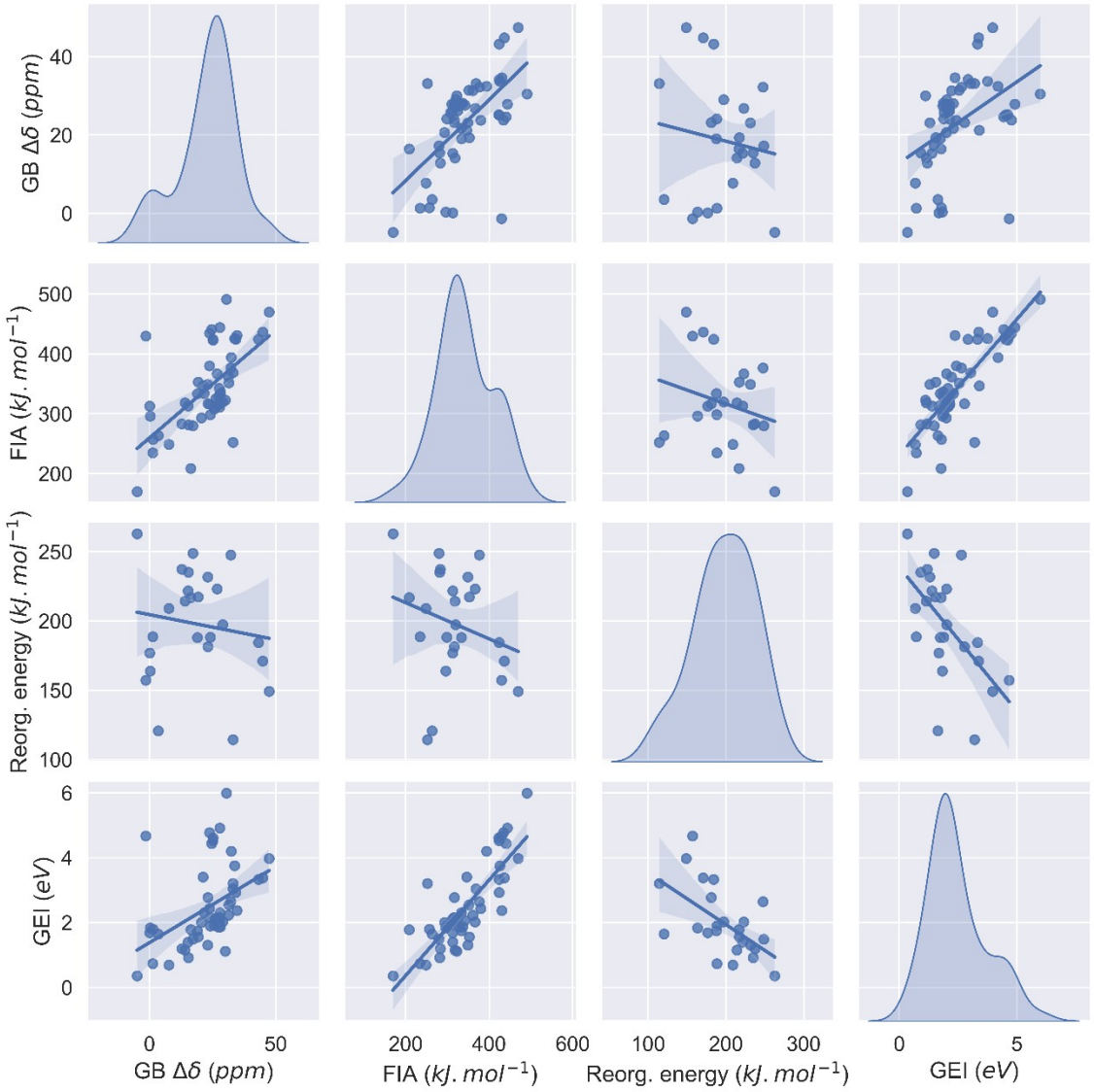


Figure S3 : pair-plot of different metrics used to quantify the Lewis acidity

	average Δδ	FIA	HIA	reorg. energy	GEI
average Δδ	1.00	0.61	0.54	-0.14	0.44
FIA	0.61	1.00	0.95	-0.27	0.76
HIA	0.54	0.95	1.00	-0.43	0.87
reorg. energy	-0.14	-0.27	-0.43	1.00	-0.47
GEI	0.44	0.76	0.87	-0.47	1.00

Table S2 : Correlation matrix (Pearson's r)

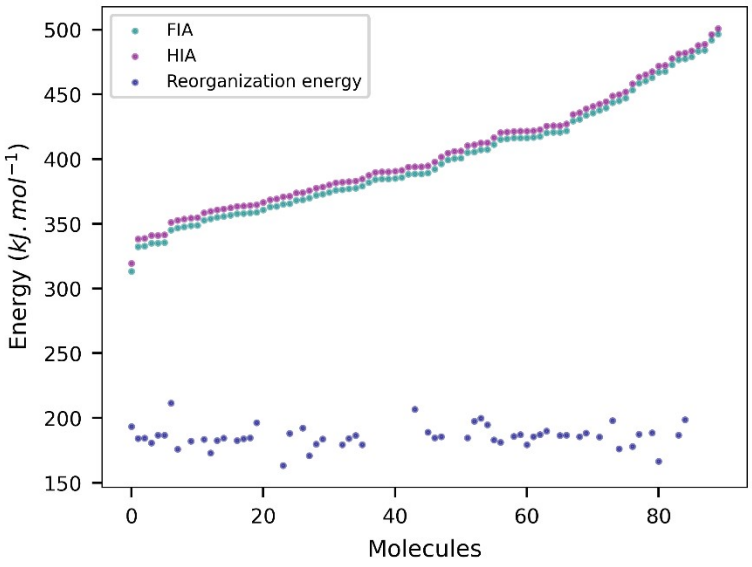


Figure S4 : FIA, HIA and reorganization energy for the ONO dataset; reorganization energy is almost constant across various functional groups substitutions on the scaffold (mean = 186 kJ.mol^{-1}, standard deviation = 8 kJ.mol^{-1}).

S2 Computational methods

FIA computation

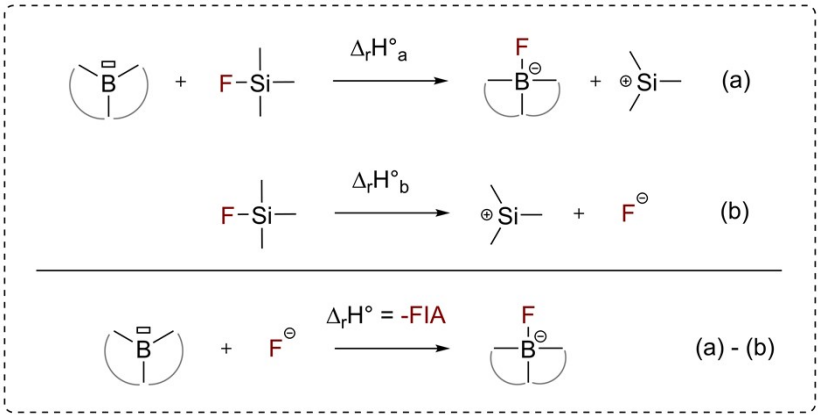
Gas phase versus solution phase calculations

Greb and colleagues have shown a strong correlation (Pearson's squared coefficient r^2 = 0.921) between FIA values calculated in the gas phase and those calculated in solution phase using dichloromethane as a solvent, based on their extensive dataset of 44k molecules.[23] This correlation is even stronger (r^2 = 0.941) for the smaller dataset of trivalent boron Lewis acids. Given that our reduced dataset also consists of boron Lewis acids with similar scaffolds, we expect an even higher correlation, allowing for a relevant comparison of FIA values in the gas phase.

Isodesmic calculations

To efficiently constitute a database of several compounds, we must employ a relatively low-cost and yet precise *ab initio* method. DFT is the typical technique employed to compute efficiently molecular properties at the *ab initio* level. However, standard DFT methods do not accurately account for electronic correlation, leading to discrepancies when compared to non-approximated *ab initio* techniques. Usually, electronic correlation errors are canceled when calculating reaction enthalpies of similar molecules. However, in case of FIA, the involvement of a "naked" fluoride ion leads to differences in the electronic correlation with respect to the molecular Lewis acid or the Lewis acid–base adduct, that are not canceled out when computing the reaction enthalpy. This is why it is recommended to employ a fluoride ion donor such as COF_3^- or Me_3SiF as a reference system and proceed in two steps: determining the reaction enthalpy between the reference system and the Lewis acid studied in the gas phase (Scheme S1, Eq. (a)), and then subtracting the dissociation enthalpy of the reference system (Scheme S1, Eq. (b)).[24,25] This approach, known as "isodesmic calculations", has been efficiently applied to reproduce FIA computed at the coupled-cluster level of theory by the Greb group.[1] The dissociation

enthalpy of the reference system must be very precise, obtained either experimentally or with a high-level *ab initio* calculation. We used the value provided by Greb et al., 952.5 kJ.mol^{-1}, obtained through CCSD(T)/CBS theory for our calculations.[1] On the other hand, the calculation of the reaction enthalpy (a) can be performed at a more flexible calculation level, as long as the precision is sufficient.



Scheme S1: Isodesmic reactions used for FIA calculation (all compounds are in the gas phase).

Choice of the DFT method

Various *ab initio* methods, mainly isodesmic, were tested on a varied set of boron-based Lewis acids, and the obtained values were compared to those tabulated by Greb et al.,[1] which were obtained at a higher level of calculation (Figure S5 and Figure S6). The DFT methods proposed by Greb et al., particularly PBEP86, converge too slowly, especially when the chosen basis set is large (aug-cc-pVTZ). Therefore, we turned to less costly methods with a MAE of the same order of magnitude (Table S1), such as M05-2X and M06-2X, along with the 6-31G(d) basis set, which had already been reported in the literature for FIA calculations.[26] M06-2X functional combined with the larger aug-cc-pVTZ basis set was also evaluated, as it was shown that M06-2X misrepresent bonding enthalpies of F-substituted complexes unless paired with a large basis set.[27] However, the isodesmic method corrects for this and there is no substantial improvement of the error with respect to the reference values by using aug-cc-pVTZ instead of 6-31G(d) (Figure S5 and S6). Besides, the computational resources (time and memory) required to perform geometry optimization and frequency calculation with the aug-cc-pVTZ basis set are too high, making impractical to efficiently compute the FIA of hundreds of molecules at this level of theory.

Regarding the linear correlations shown in Figure S6, ideally, the slope should be 1 (which is roughly verified) and the intercept should be 0. However, a non-zero intercept only introduces a systematic bias and still allows us to observe trends, which allows the relative comparison of Lewis acids with one another.

Since the performance of the M06-2X method was slightly superior to that of the M05-2X method (refer to the MAE values in Table S3), and the use of larger basis set too computationally demanding, we chose M06-2X combined with the 6-31G(d) basis set, in isodesmic calculations, to generate the FIA data needed for database construction.

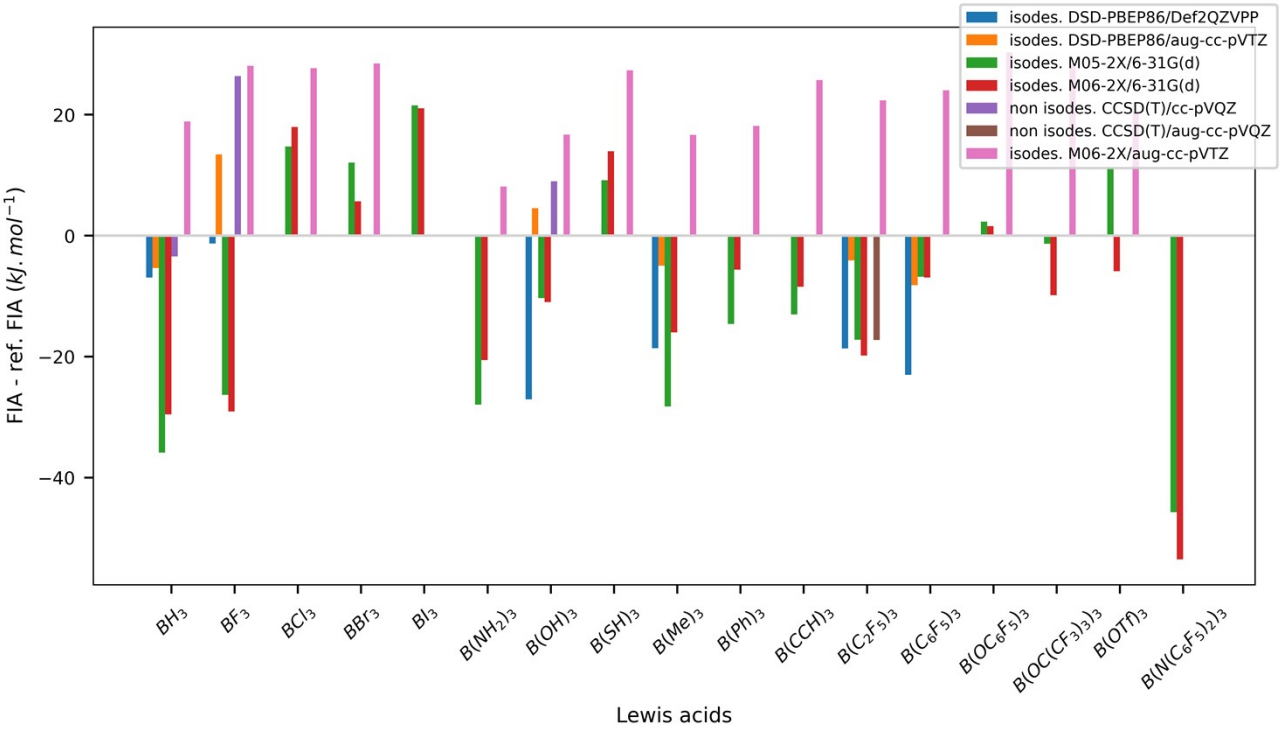


Figure S5 : FIA values relative to the reference values for each species and method.

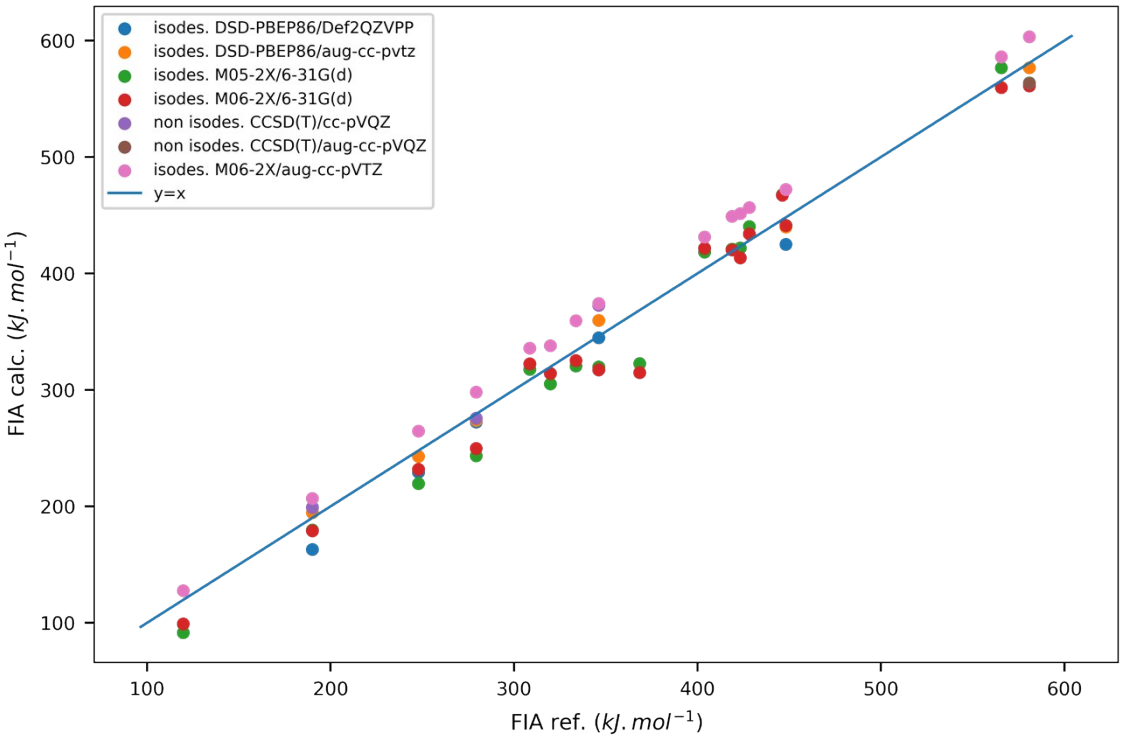


Figure S6 : Linear correlation between the calculated FIA values and the reference values.

Method	isodes. DSD-PBEP86/def2QZVPP	isodes. DSD-PBEP86/aug-cc-pVTZ	isodes. M05-2X/6-31G(d)	isodes. M06-2X/6-31G(d)	Isodes. M06-2X/aug-cc-pVTZ
Slope	1	0.98	1.07	1.03	1.03
Intercept	-16.07	4.74	-36.62	-21.77	12.58
R^2	0.995	0.997	0.982	0.979	1
MAE	7.89	5.78	13.42	13.1	4.36

Table S3 : parameters of the linear correlations plotted in Figure S6 and MAE

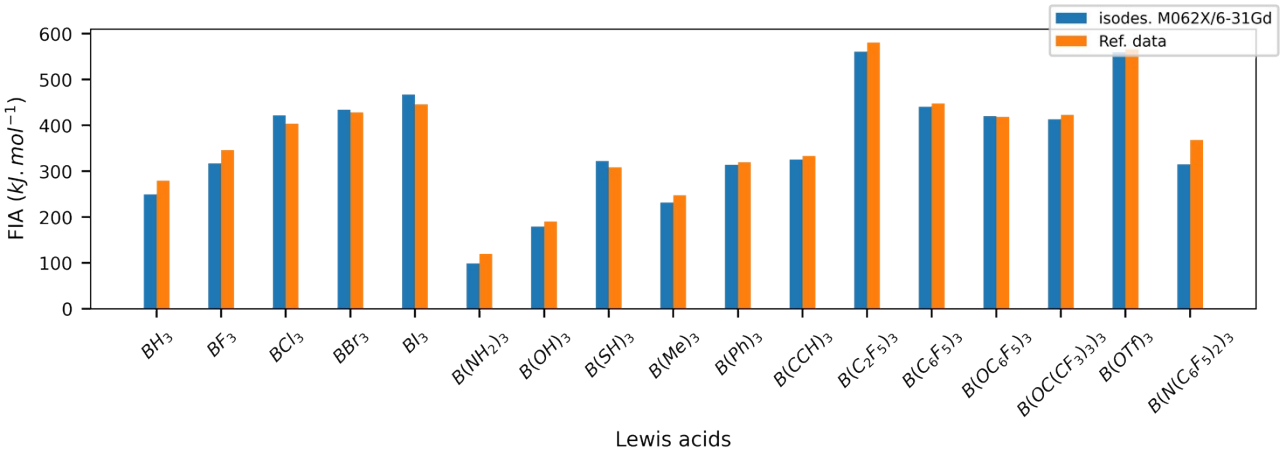


Figure S7 : FIA values calculated (M06-2X/6-31G(d)) and reference FIA values (from higher-level calculations) found in the literature for a set of boron-based Lewis acids.

HIA computation

HIA data were computed primarily for comparison with FIA. For a given molecular scaffold, FIA and HIA are highly correlated (Figure S4). Unlike FIA, HIA was computed without using isodesmic calculations.

GEI computation

The global electrophilicity index (GEI) was calculated from frontier molecular orbital energies (E_{HOMO} and E_{LUMO}) obtained at the M06-2X/6-31G(d) level of theory with:

$$GEI = \frac{\chi^2}{2\eta}$$

Where, the global electronegativity can be calculated as $\chi = \left(\frac{E_{HOMO} + E_{LUMO}}{2}\right)$ and the hardness can be calculated as $\eta = (E_{LUMO} - E_{HOMO})$.

Reorganization energy computation

Reorganization energy calculations are conceptually trivial. The energy of reorganization is calculated in the following way: the fluoride ion of the borate is removed and a single point energy (SPE) calculation is performed on the boron derivative bended geometry obtained. The energy of the Lewis acid with optimized geometry is subtracted to this quantity to yield the reorganization energy. Yet, as lots of SPE calculations failed (blanks in table S1), we computed less reorganization energies compared to the other metrics and did not pursue in computing missing values.

$$reorg.energy = E(LA\ in\ the\ adduct\ conformation) - E\ (LA\ in\ optimized\ conformation)$$

S3 Chemical space

Database extension

To the prediction performances of the models, we aimed to extend the database to at least 200 molecules. To achieve this, we performed a database extension for the ONO scaffold using the K-means[28] and coverage[29] clustering algorithms, targeting 50 clusters for each algorithm. These algorithms were applied to a Morgan fingerprint representation of 2,197 possible molecules in the chemical space, with the constraint of excluding molecules already present in the database. The K-means algorithm was repeated 10 times, while the coverage algorithm, that requires a longer convergence time, was repeated 4 times. From these runs, we selected the 50 molecules that appeared most frequently across all repetitions of both algorithms. Figure S8 shows the Tanimoto Similarity distribution of the selected molecules, which displays a relatively broad Gaussian distribution centered around a similarity of 0.35. This similarity value suggests a satisfying diversity in the selected molecular structures, indicating that these clustering algorithms effectively sampled the chemical space in a uniform manner. In total, we obtained 97 unique molecules (as 3 molecules were common to both algorithms), for which we calculated the FIA. As the K-means clustering algorithm was faster to process, we preferred it to extend the database of the triarylboranes of 100 molecules, for molecular design purposes.

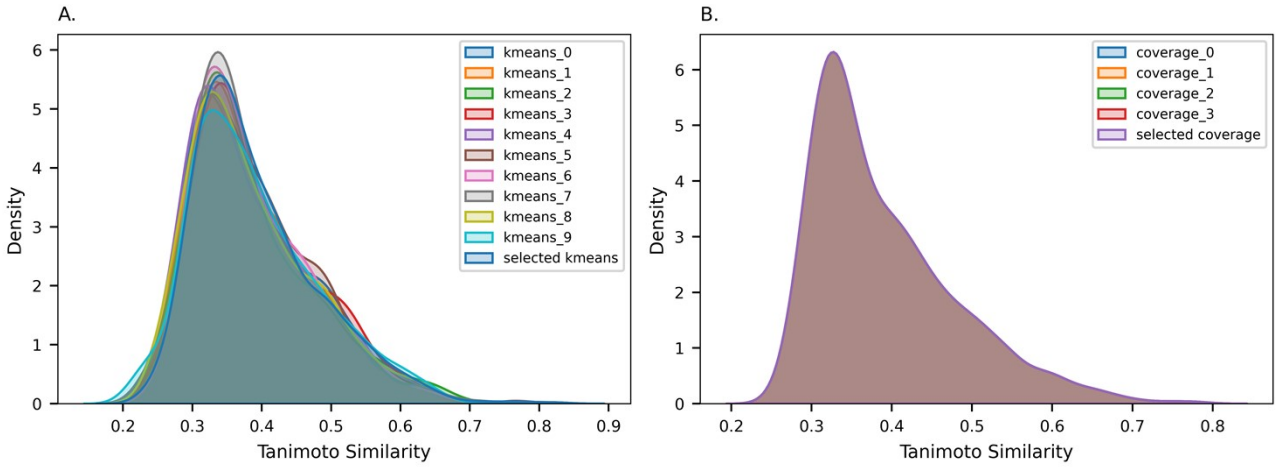


Figure S8 : Tanimoto similarity density curves of selected molecules for database extension for ONO scaffold, A. results for the K-means clustering algorithm, B. results for the coverage clustering algorithm.

Database curation

FIA values were computed for each molecule and compiled into tables where molecules are represented as SMILES[30] strings. Outliers were detected plotting the distribution of FIA values and conducting PCA on quantum descriptors. These outliers were then eliminated from the database.

S4 Molecular descriptors

Chemo-informatics descriptors

We used the extended connectivity Morgan Fingerprints[31] and RDKit descriptors.[32,33] Fingerprints are 1024 bit-vectors emphasizing molecular connectivity and fragment diversity. RDKit descriptors provide structural and property-based features, including fragment counts, minimum and maximum absolute partial charges, molecular weight, and other QSAR descriptors (208 features in total).

Quantum descriptors

Quantum features for each molecule were derived from the DFT calculation performed for FIA computation, using the Auto-QChem[34] workflow developed by the Doyle Group. This enabled efficient extraction of precise and physically meaningful features (amounting to 43), including global molecular properties (e.g., frontier orbitals energies) as well as local atomic properties for the boron atom only (e.g., partial charge). In spite of requiring more computational resources to compute, these descriptors offer reliable insights into the relationship between the quantum features of molecules and their Lewis acidity (see the manuscript, *Interpretability* section).

Hammett-extended descriptors

These descriptors were generated by identifying the chemical nature of substituents at *ortho*, *meta* and *para* positions of each scaffold using the SMARTS substructure identifiers[35] implemented in the RDKit python library.[32] The code to generate the descriptors is available on the GitHub repository of the project.[36] Then, Hammett-extended parameters derived by Sigman and co-workers[37] corresponding to the *ortho*, *meta* and *para* substituents were concatenated into a vector featuring the molecule (36 features). These parameters, characteristic of the substituents on the benzoic acid, are the Sterimol parameters, *B1, B5* and *L* (in Å), infrared (IR) spectroscopy frequencies, *v* (in cm^{-1}) and intensities *I*, Natural Bonding Orbital (NBO) charges (in e, elementary charge), and torsional angle between the aromatic ring plane and the carboxylic moiety, ϑ_{tor} (in °) (see Table S4, Table S5, and Table S6).

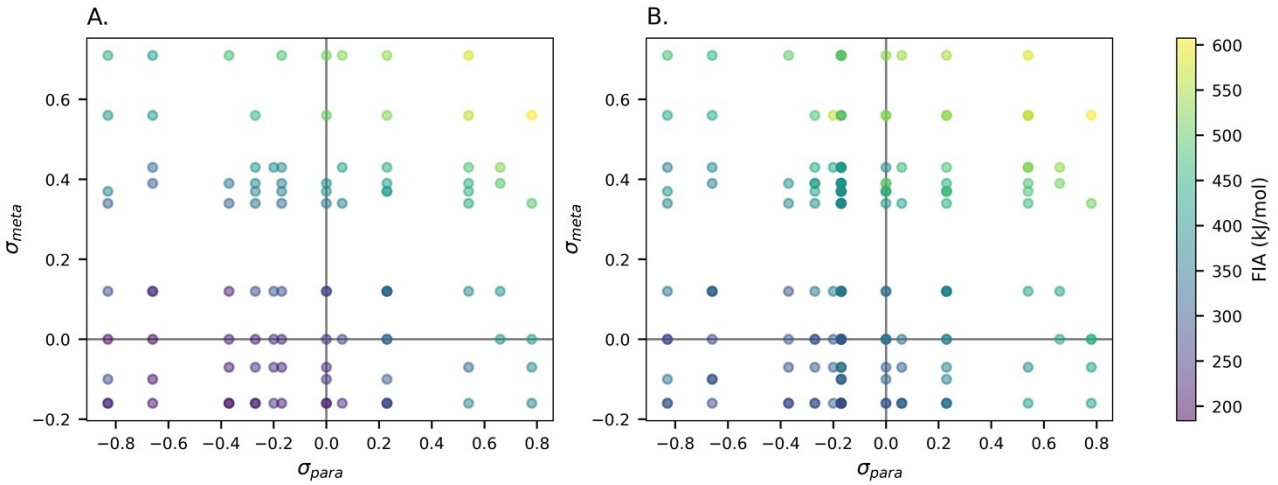


Figure S9 : Representations of the triarylboranes dataset following the value of their *para* and *meta* σ constants. A. Molecules from the triaryboranes dataset with no substituent in the *ortho* position. B. All molecules from the triarylborane dataset.

R-ortho	$B_{1,o}$	$B_{5,o}$	L_o	$\nu_{COH,o}$	$I_{COH,o}$	$\nu_{C=O,o}$	$I_{C=O,o}$	$NBO_{C,o}$	$NBO_{=O,o}$	$NBO_{O,o}$	$NBO_{H,o}$	ϑ_{tor}
H	1.00	1.00	2.09	1394.60	163.25	1847.85	390.68	0.80516	-0.60208	-0.70358	0.50087	0.00
NMe_2	1.55	3.27	4.30	1385.34	46.99	1826.81	455.50	0.80697	-0.61279	-0.69946	0.49588	25.77
NH_2	1.55	2.06	3.03	1388.97	76.63	1830.53	481.33	0.80849	-0.61276	-0.73355	0.50600	2.60
OH	1.52	2.07	2.83	1408.63	122.13	1825.36	430.86	0.80852	-0.61647	-0.68045	0.49706	0.00
OCH3	1.52	3.22	4.25	1399.72	131.89	1837.11	429.50	0.81217	-0.60468	-0.69401	0.49943	21.20
***t*Bu**	2.91	3.35	4.34	1378.50	99.39	1852.54	355.57	0.82059	-0.59468	-0.70647	0.49794	56.76
Me	1.70	2.20	3.07	1391.63	122.69	1831.67	408.67	0.80753	-0.61086	-0.70423	0.50162	0.00
F	1.47	1.47	2.80	1407.80	155.11	1837.86	410.90	0.80809	-0.60458	-0.68210	0.49971	0.00
Cl	1.77	1.77	3.47	1403.56	136.98	1838.98	404.79	0.81116	-0.59957	-0.68514	0.50130	25.80
Br	1.92	1.92	3.77	1401.92	130.52	1842.55	407.66	0.81188	-0.59658	-0.68619	0.50127	30.30
CF_3	2.08	2.71	3.58	1402.74	230.78	1854.51	371.41	0.80925	-0.58821	-0.68729	0.50302	36.60
CN	1.78	1.78	4.18	1403.17	160.60	1847.07	386.94	0.80262	-0.59353	-0.68333	0.50810	29.19
NO_2	1.55	2.58	3.61	1395.67	112.67	1869.31	375.43	0.81941	-0.57682	-0.69015	0.50473	44.26

Table S4 : Hammett-extended parameters for the *ortho* substituent.[37]

R-meta	$B_{1,m}$	$B_{5,m}$	***Lm***	σ_m	$\nu_{COH,m}$	$I_{COH,m}$	$\nu_{C=O,m}$	$I_{C=O,m}$	$NBO_{C,m}$	$NBO_{=O,m}$	$NBO_{O,m}$	$NBO_{H,m}$
H	1.00	1.00	2.09	0.00	1394.60	163.25	1847.85	390.68	0.80516	-0.60208	-0.70358	0.50087
NMe_2	2.00	3.28	4.11	-0.16	1378.42	106.94	1845.61	398.56	0.80655	-0.60340	-0.70782	0.49914
NH_2	1.55	2.08	2.99	-0.16	1398.20	172.81	1847.74	379.29	0.80639	-0.60132	-0.70622	0.49971
OH	1.52	2.07	2.86	0.12	1398.66	189.34	1849.36	370.78	0.80529	-0.59959	-0.70183	0.50059
OCH3	1.52	3.20	4.25	0.12	1396.67	216.86	1847.51	385.34	0.80599	-0.60040	-0.70218	0.50019
***t*Bu**	2.92	3.35	4.36	-0.10	1395.11	148.49	1845.19	426.62	0.80583	-0.60299	-0.70606	0.50005
Me	1.70	2.19	3.07	-0.07	1395.49	157.59	1846.08	400.44	0.80602	-0.60256	-0.70452	0.49996
F	1.47	1.47	2.81	0.34	1397.94	186.91	1852.75	369.96	0.80394	-0.59554	-0.70171	0.50225
Cl	1.77	1.77	3.51	0.37	1394.82	180.38	1852.53	389.42	0.80547	-0.59554	-0.70118	0.50252
Br	1.92	1.92	3.81	0.39	1394.08	177.79	1851.09	396.40	0.80558	-0.59541	-0.70094	0.50265
CF_3	2.09	2.72	3.48	0.43	1408.57	201.35	1853.23	388.18	0.80364	-0.59415	-0.70028	0.50352

CN	1.78	1.78	4.18	0.56	1401.10	182.03	1856.38	395.52	0.80490	-0.59114	-0.69879	0.50472
NO_2	1.55	2.59	3.56	0.71	1399.96	179.66	1856.20	387.73	0.80328	-0.59110	-0.69693	0.50490

Table S5 : Hammett-extended parameters for the *meta* substituent.[37]

R-para	$B_{1,p}$	$B_{5,p}$	L_p	σ_p	$\nu_{COH,p}$	$I_{COH,p}$	$\nu_{C=O,p}$	$I_{C=O,p}$	$NBO_{C,p}$	$NBO_{=O,p}$	$NBO_{O,p}$	$NBO_{H,p}$
H	1.00	1.00	2.09	0.00	1394.60	163.25	1847.85	390.68	0.80516	-0.60208	-0.70358	0.50087
NMe_2	1.55	3.30	4.33	-0.83	1396.43	313.71	1830.85	486.47	0.80471	-0.62021	-0.71045	0.49757
NH_2	1.55	2.10	2.98	-0.66	1395.35	257.80	1834.97	453.68	0.80524	-0.61626	-0.70910	0.49865
OH	1.52	2.07	2.83	-0.37	1396.61	221.99	1840.65	422.67	0.80599	-0.61060	-0.70602	0.50016
OCH3	1.52	3.22	4.25	-0.27	1396.68	299.76	1839.37	424.99	0.80480	-0.61182	-0.70667	0.49976
***t*Bu**	2.91	3.35	4.34	-0.20	1395.78	189.95	1844.67	435.29	0.80613	-0.60531	-0.70473	0.49999
Me	1.70	2.20	3.07	-0.17	1395.02	178.02	1844.66	419.34	0.78588	-0.60618	-0.70599	0.50033
F	1.47	1.47	2.80	0.06	1396.19	204.97	1847.75	398.96	0.80562	-0.60242	-0.70414	0.50193
Cl	1.77	1.77	3.47	0.23	1395.38	186.22	1849.16	414.28	0.80492	-0.59926	-0.70285	0.50238
Br	1.92	1.92	3.77	0.23	1395.33	182.92	1849.58	419.46	0.80491	-0.59850	-0.70255	0.50251
CF_3	2.08	2.71	3.58	0.54	1401.11	45.48	1854.66	375.38	0.78303	-0.59324	-0.70152	0.50390
CN	1.78	1.78	4.18	0.66	1397.32	181.63	1855.52	393.01	0.80177	-0.58965	-0.69956	0.50474
NO_2	1.55	2.58	3.61	0.78	1396.61	202.70	1857.56	370.43	0.78101	-0.58788	-0.69977	0.50541

Table S6 : Hammett-extended parameters for the *para* substituent.[37]

Hammett-extended descriptors cannot account for differences in FIA values for molecules exhibiting the same substituents but with different molecular scaffolds (Figure S10).

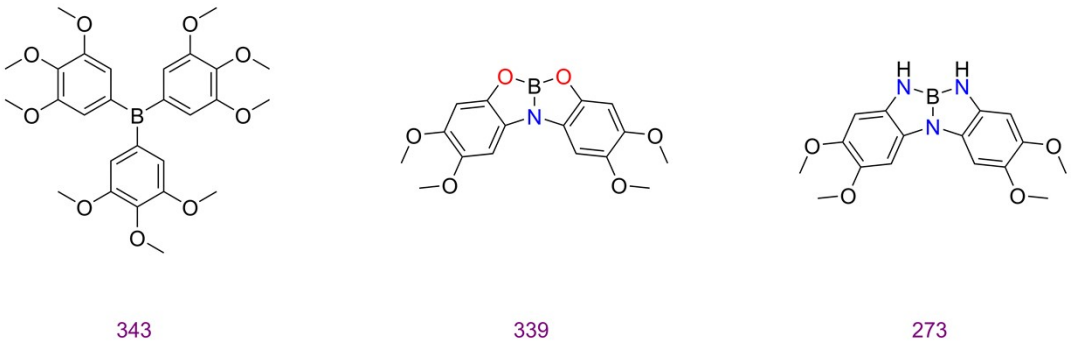


Figure S10 : Varying FIA values for molecules with the same substituents but differing molecular scaffolds.

S5 Constructing machine learning models

Dataset splitting

To address the low-data regime, stratified sampling was used to split the dataset into training and testing sets, preserving the FIA distribution across both. Classes defined in the **Interpretability**, *Molecular Design* – Interpretable ML section of the manuscript for decision tree fitting were used for stratification, ensuring the testing set was representative of the overall FIA population for the ONO scaffold (Figure S11).

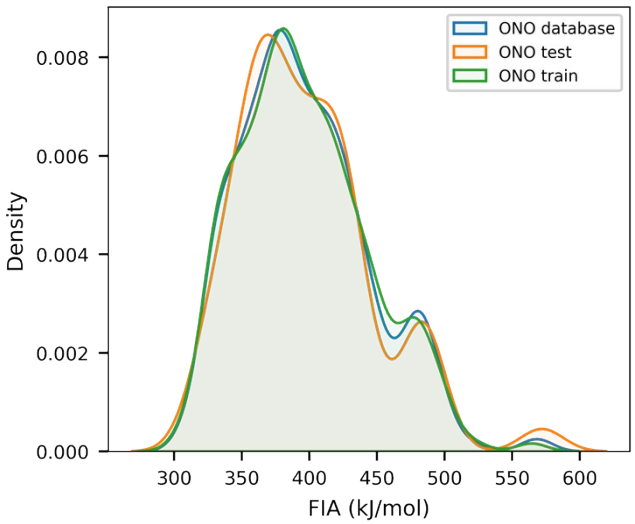


Figure S11 : FIA distribution for the whole database, for training and for testing set defined for model performances evaluation.

Models' evaluation

Data preprocessing

Before any learning (including folds of cross validation), data was preprocessed by removing constant features, fitting a standard scaler on the training set and applying it on both the training and validation sets. This way, all features were scaled to have 0 mean and unit standard deviation. This scaling process was not applied to Morgan fingerprints, as they are bit-vectors. For fingerprints, only constant features across datapoint molecules (0 variance) were removed.

The mean absolute error (MAE) was chosen as the scoring function to evaluate the models, as it provided the advantage to possess a unit, which makes it easier to figure out what physically means the error in a regression task. Models were evaluated and optimized on the training set. The hyperparameters of the models were tuned beginning with a randomized search and then with a grid search 4-fold cross-validation scheme. Final tuning of the hyperparameters was performed manually using a 10-fold cross-validation scheme. That scheme was repeated 10 times to statistically evaluate the distribution of errors for each model.

The obtained models were used to conduct the rest of this study. Chosen hyperparameters are available in the GitHub repository.[36]

Oracle development

Linear regression (LR) was chosen as the ML algorithm due to its strong performance across a broad range of molecular descriptors. Concatenated features from the Hammett-extended and RDKit descriptors were filtered using Scikit-learn's **SelectKBest** feature selection algorithm with **f_regression** as the scoring function.[38] This function evaluates features based on their F-statistic, measuring the strength of their linear relationship with the target FIA when considered independently. This approach retained only the most relevant features for the regression task, with the optimal number of features determined through a grid search involving also the model's hyperparameters to minimize the MAE. Notably,

features such as structural fragment descriptors from RDKit and Sterimol parameters like B_1 and B_5 from the Hammett-extended descriptors were excluded due to their low correlation with FIA. See the GitHub repository for details.[36]

Extrapolation

Quantum descriptors

Linear	LR	Bayes. Ridge	LASSO	SVR	Tree	RF	Grad. Boost.	GPR	KNN	MLP
MAE = 54.1	MAE = 187.8	MAE = 227.3	MAE = 163.6	MAE = 86.5	MAE = 76.3	MAE = 71.2	MAE = 61.9	MAE = 86.4	MAE = 60.2	MAE = 24.1
a = 0.89	a = 0.88	a = 0.88	a = 0.88	a = 0.86	a = 0.81	a = 0.84	a = 0.76	a = 0.0	a = 0.45	a = 0.89
b = -18.8	b = 226.6	b = 266.5	b = 200.9	b = 130.1	b = 135.1	b = 121.5	b = 137.3	b = 396.7	b = 231.3	b = 13.5

Table S7 : Performances of various ML algorithms trained on the ONO dataset and tested on the NNN dataset with the quantum descriptors.

	ONO	NNN	OCO	Triarylboranes
ONO	MAE = 8.13 a = 0.95 b = 19.03	MAE = 692.9 a = 0.78 b = 780.95	MAE = 104.03 a = 0.86 b = 158.01	MAE = 268.7 a = 1.41 b = -430.62
NNN	MAE = 187.79 a = 0.88 b = 226.62	MAE = 9.56 a = 0.95 b = 16.76	MAE = 283.4 a = 0.88 b = -244.77	MAE = 164.47 a = 1.4 b = -290.21
OCO	MAE = 35.09 a = 0.92 b = -8.08	MAE = 493.86 a = 0.76 b = 574.82	MAE = 11.21 a = 0.89 b = 39.83	MAE = 114.52 a = 0.99 b = -111.94
triarylboranes	MAE = 728.52 a = 0.93 b = -696.08	MAE = 378.86 a = 0.38 b = -135.1	MAE = 417.73 a = 0.31 b = -140.58	MAE = 15.85 a = 0.92 b = 30.67

Table S8 : MAE and fit coefficients for training scaffolds (vertical) and testing scaffolds (horizontal) for the LR regressor with the quantum descriptors (no feature selection). MAE obtained for prediction within the same scaffold were averaged over a 10-fold cross-validation scheme, repeated 5 times.

	NNN	ONO	deviation
X	0.000	-0.001	1.424
lumo_energy	-0.002	-0.010	1.239
Z	-0.010	-0.003	1.031
NPA_Rydberg	0.016	0.029	0.575
ES_root_NPA_Rydberg	0.017	0.030	0.564
NMR_anisotropy	16.345	11.588	0.341
dipole	4.841	3.584	0.298
Y	-0.538	-0.689	0.247
ES_root_NPA_charge	1.045	1.281	0.203
NPA_charge	1.048	1.258	0.182
APT_charge	1.067	1.254	0.161
ES_root_dipole	4.674	3.996	0.156

ES_root_NPA_valence	1.940	1.691	0.137
NPA_valence	1.937	1.714	0.122
G_thermal_correction	0.255	0.231	0.1
zero_point_correction	0.308	0.283	0.086
E_thermal_correction	0.331	0.305	0.083
H_thermal_correction	0.332	0.306	0.083
electronegativity	0.127	0.138	0.08
ES_root_NPA_total	3.955	3.719	0.062
number_of_atoms	40.918	38.485	0.061
NPA_total	3.952	3.742	0.055
homo_energy	-0.252	-0.265	0.051
H	-2440.757	-2541.185	0.04
electronic_spatial_extent	13522.813	12986.665	0.04
E_scf	-2440.941	-2541.330	0.04
E_zpe	-2440.781	-2541.208	0.04
G	-2440.833	-2541.260	0.04
E	-2440.758	-2541.186	0.04
NMR_shift	89.842	86.465	0.038
ES_root_molar_volume	2671.401	2588.301	0.032
molar_volume	2688.979	2605.690	0.031
ES_root_electronic_spatial_extent	13768.427	13392.656	0.028
VBur	0.605	0.590	0.025
hardness	0.125	0.127	0.021
ES_root_Mulliken_charge	0.670	0.683	0.02
molar_mass	395.729	390.020	0.015
Mulliken_charge	0.668	0.667	0.001
converged	1.000	1.000	0
ES_root_NPA_core	1.999	1.999	0
NPA_core	1.999	1.999	0
multiplicity	1.000	1.000	0

Table S9 : mean of quantum features for ONO and NNN scaffolds, and the deviation between the two scaffolds. This deviation was calculated by the formula: $deviation_{feature} = (mean_{feat.,\ NNN}) - mean_{feat.,ONO})/(mean_{feat.,NNN} + mean_{feat.,ONO})$.

features	corr with FIA
homo_energy	0.749
electronegativity	0.704
lumo_energy	0.628
G_thermal_correction	0.448
zero_point_correction	0.403
G	0.378
H	0.378
E	0.378
E_zpe	0.378
E_scf	0.378
H_thermal_correction	0.375
E_thermal_correction	0.375

NPA_total	0.326
NPA_charge	0.326
NPA_valence	0.324
molar_mass	0.306
Mulliken_charge	0.277
ES_root_electronic_spatial_extent	0.271
electronic_spatial_extent	0.27
ES_root_Mulliken_charge	0.237
ES_root_NPA_total	0.236
ES_root_NPA_charge	0.236
ES_root_NPA_valence	0.233
number_of_atoms	0.23
VBur	0.164
APT_charge	0.158
NMR_anisotropy	0.129
dipole	0.127
NMR_shift	0.092
ES_root_NPA_Rydberg	0.09
NPA_Rydberg	0.076
hardness	0.07
ES_root_dipole	0.057
Z	0.054
NPA_core	0.039
ES_root_NPA_core	0.039
Y	0.031
molar_volume	0.022
X	0.009
ES_root_molar_volume	0.004
charge	
converged	
multiplicity	

Table S10 : correlation of quantum features with FIA (Pearson's correlation coefficient), total chemical space (ONO, NNN, OCO and triarylboranes).

Details of feature selection

The selected quantum features for the prediction from the ONO scaffold to the NNN scaffold are : E, ES_root_electronic_spatial_extent, ES_root_molar_volume, E_scf, E_thermal_correction, E_zpe, G, G_thermal_correction, H, H_thermal_correction, electronegativity, electronic_spatial_extent, hardness, homo_energy, molar_mass, molar_volume, number_of_atoms, zero_point_correction, APT_charge, ES_root_Mulliken_charge, ES_root_NPA_charge, ES_root_NPA_core, ES_root_NPA_total, ES_root_NPA_valence, Mulliken_charge, NMR_anisotropy, NMR_shift, NPA_charge, NPA_core, NPA_total, NPA_valence.

Exploration of the inherent differences between scaffolds

Even when using all quantum features, models exhibited different biases depending on the structure (Table S8). This variability may stem from inherent differences between molecular scaffolds, not only in terms of quantum features but

also in FIA distribution (Figure 2.B). Indeed, most ML algorithms perform well when the distribution of the training set resembles that of the testing set. Expecting the model would capture the differences between scaffolds, we trained it on three molecular scaffolds and then tested on the fourth. This approach resulted in improved performances on ONO and NNN (MAE decreased of 50 to 100 kJ.mol^{-1}), but not on the other structures (Table S11). The low prediction performance on triarylboranes is expected (MAE = 466 kJ.mol^{-1}), given their structural differences from the others, being non planar and lacking heteroatoms bonded to the boron atom (Figure 6). Furthermore, while the FIA range for the OCO structure falls within the overlap of distributions for the other structures, its FIA distribution significantly differs due to its sharpness, possibly explaining the low prediction performance (MAE = 157 kJ.mol^{-1}). This sharp distribution may also disrupt FIA learning when predicting on other structures. Indeed, excluding the OCO structure significantly improved performances on ONO (MAE = 25.2 kJ.mol^{-1}) and NNN (MAE = 12.87 kJ.mol^{-1}) (Table S12).

	ONO	**NNN**	**OCO**	**Triarylboranes**
Performance	MAE = 73.23	MAE = 41.13	MAE = 157.11	MAE = 466.06
	a = 1.11	a = 1.05	a = 0.85	a = 0.79
	b = -117.61	b = 24.72	b = 209.43	b = -382.58

Table S11 : Performances of LR and quantum descriptors model, tested on one structure while trained on the three others.

	ONO	**NNN**	**Triarylboranes**
Performance	MAE = 25.21	MAE = 12.87	MAE = 576.75
	a = 1.18	a = 0.95	a = 0.78
	b = -94.64	b = 21.21	b = -490.18

Table S12 : Performances of LR and quantum descriptors model, tested on one structure while trained on the two others (OCO structure excluded).

	ONO	**NNN**	**OCO**
Performance	MAE = 42.54	MAE = 83.5	MAE = 45.28
	a = 0.92	a = 0.9	a = 0.91
	b = -11.73	b = -52.42	b = -13.14

Table S13 : Performances of LR and quantum descriptors model, tested on one structure while trained on the two others (triarylboranes excluded).

RDKit descriptors

Given the complexity of the RDKit descriptors, we implemented a slightly different strategy for extrapolating. We again selected the LR algorithm, as it demonstrates strong performance, achieving a MAE of 6.93 kJ.mol^{-1} and 10.5 kJ.mol^{-1} for the ONO and NNN respectively (Table S14). And again, when trying to predict on the NNN while trained on the ONO, the MAE of the model significantly increases to 1828 kJ.mol^{-1}. We then used a recursive feature elimination strategy, by iteratively removing features when it improved model predictions from ONO to NNN. Removing FractionCSP3, BCUT2D_CHGHI, and MinEStateInde really improved the performance of the model (MAE dropped to 19.97 kJ.mol^{-1}). Eliminating SlogP_VSA2, the MAE could be reduced to 9.05 kJ.mol^{-1},while removing only 4 features out of 208. Prediction

performance remained high within the ONO (7.11 kJ.mol^{-1}) and the NNN (10.6 kJ.mol^{-1}) chemical spaces using these selected features. Results are illustrated in Figure S12.

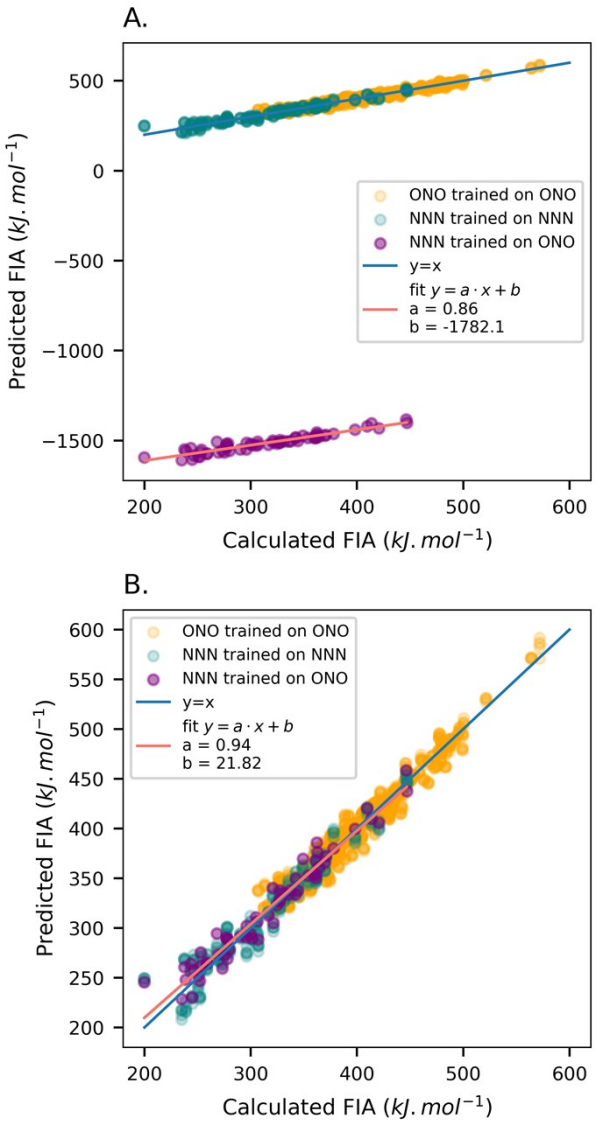


Figure S12 : Feature selection to extrapolate from ONO to NNN with the LR regressor and RDKit descriptors. A. No feature selection. B. Features selected.

Despite these improvements, when attempting to extrapolate to other molecular scaffolds using the same selected features, the MAE remains considerably large (267.77 kJ.mol^{-1} for OCO and 384.78 kJ.mol^{-1} for triarylboranes), although notably reduced with respect to the initial values (Table S14). Training on three scaffolds and testing on the remaining one while keeping all RDKit features significantly improved performances (Table S15). This strategy, combined with the selected RDKit features, reduced MAE of 10 to 30 kJ.mol^{-1} (Table S16).

	ONO	**NNN**	**OCO**	**triarylboranes**
ONO	MAE = 6.93 a = 0.97 b = 13.6	MAE = 4264.81 a = 1.0 b = 4263.01	MAE = 3192.32 a = 0.61 b = 3345.35	MAE = 126.18 a = 0.74 b = 229.8

NNN	MAE = 1827.67 a = 0.86 b = -1782.21	MAE = 10.46 a = 0.95 b = 15.98	MAE = 3990.9 a = 0.56 b = -3851.26	MAE = 141.15 a = 0.62 b = 259.66
OCO	MAE = 1011.51 a = 1.04 b = -1023.54	MAE = 2209.3 a = 0.97 b = 2218.55	MAE = 9.8 a = 0.89 b = 36.66	MAE = 130.61 a = 0.88 b = 173.01
Triarylboranes	MAE = 3477.93 a = 0.68 b = -3354.13	MAE = 4909.6 a = 9.65 b = -8256.87	MAE = 9326.99 a = -0.73 b = -8656.36	MAE = 28.94 a = 0.81 b = 74.53

Table S14 : MAE and fit coefficients for training scaffolds (vertical) and testing scaffolds (horizontal) for the LR regressor with non-selected RDKit features model (10-fold cross-validation scheme).

	ONO	**NNN**	**OCO**	**Triarylboranes**
Performance	MAE = 37.7 a = 0.81 b = 37.66	MAE = 56.25 a = 0.72 b = 145.36	MAE = 56.42 a = 0.98 b = 62.64	MAE = 87.41 a = 1.08 b = -93.03

Table S15 : Performances of LR and non-selected RDKit descriptors model, tested on one structure while trained on the three others.

	ONO	**NNN**	**OCO**	**Triarylboranes**
Performance	MAE = 21.49 a = 0.77 b = 72.2	MAE = 45.33 a = 0.73 b = 131.67	MAE = 25.06 a = 0.98 b = 28.95	MAE = 79.38 a = 1.11 b = -88.95

Table S16 : Performances of LR and selected RDKit descriptors model, tested on one structure while trained on the three others.

Hammett-extended descriptors

As expected for the Hammett-extended descriptors that do not take into account the molecular scaffold, no feature selection could improve the MAE of 86.4 kJ.mol^{-1} to predict from ONO to NNN.

S6 Interpretability

Lewis acidity interpretability

Principal component analysis (PCA)

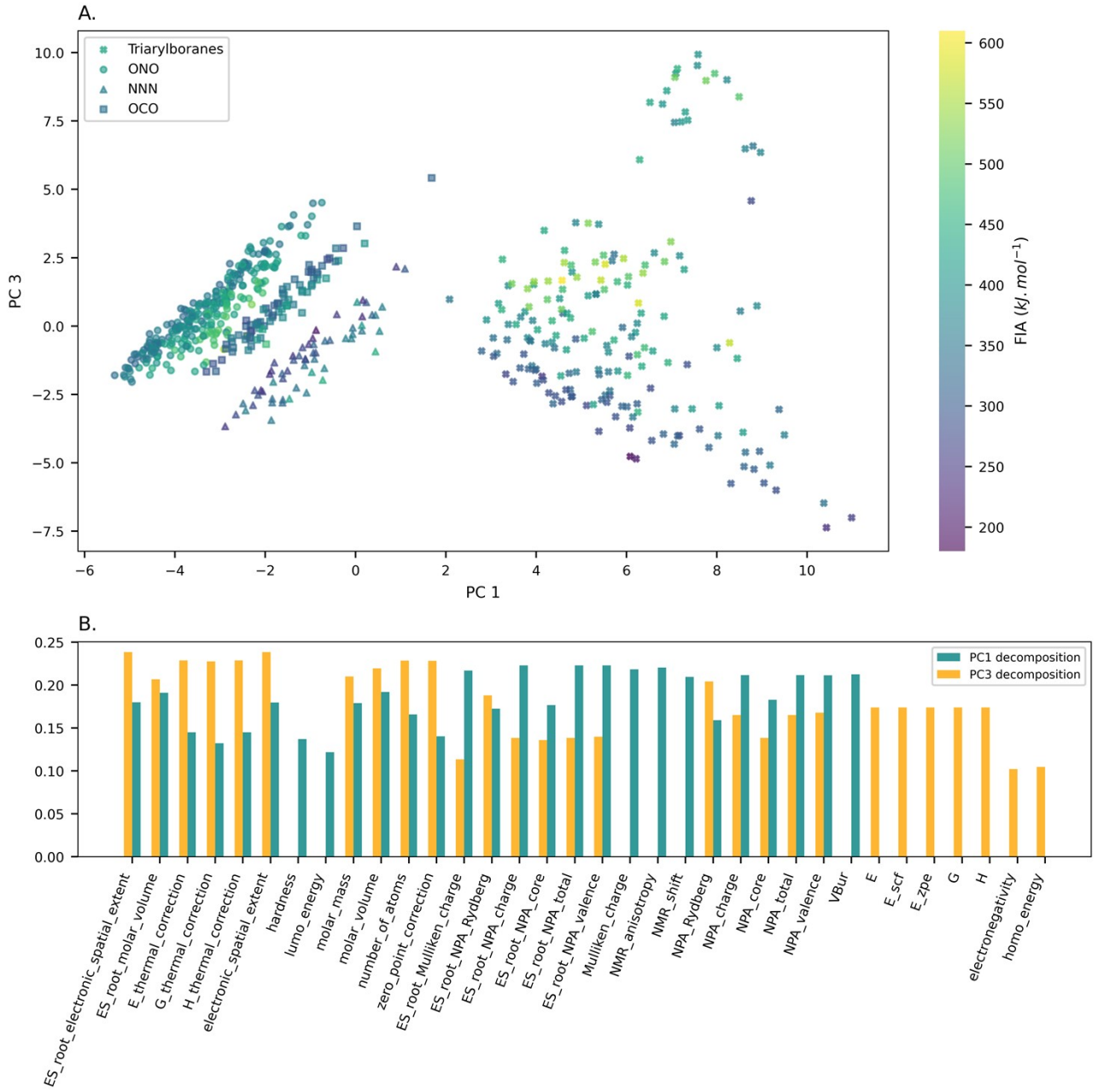


Figure S13: Principal component analysis (PCA) of the quantum descriptors. A. Projection onto the first and the third principal components. Molecules are colored according to their FIA values. B. Explained variance ratio for each component. C. Contributions of the quantum descriptors features to the first two principal components.

In addition to the PCA projection onto the first two principal components (PC1 & PC2) shown in the manuscript (Figure 6), we also examined the projection onto PC1 and PC3 (Figure S13). This revealed a separation of the constrained scaffold molecules cluster present in Figure 6, primarily along the PC3 axis. PC3 is partially decomposed along molecular scale

features such as molar mass, volume and absolute energy features E, G and H), which are characteristic of certain scaffolds. Therefore, this observed separation is unsurprising.

Decorrelating features

Various feature interpretation techniques presuppose that the features are not correlated, which is why we first searched to decorrelate features prior to any analysis. We computed the correlation matrix (using the Spearman correlation coefficient) and performed hierarchical clustering to identify uncorrelated features. The threshold to separate clusters was set to 0.45 (Figure S13.A). Then we chose one feature by cluster (the easiest to interpret physically), except for boron atom coordinates for which we chose only the X coordinate. Even if coordinate features are not expected to play a role in predicting FIA (initial geometries are generated without standardization), we chose to keep one representative (the X coordinate) to prove it. Quantitative confirmation that this feature has no influence on FIA was obtained in the further analyses (permutation importance, Figure S13.B and linear model coefficients, Figure S16). The decorrelation analysis resulted in 12 chosen uncorrelated features (absolute enthalpy *H*, dipolar moment, molar volume, thermal correction to enthalpy, global electronegativity, hardness, Atomic Polar Tensor charge of boron atom, natural charge of boron atom from Natural Population Analysis, number of electrons in the Rydberg orbitals of boron, number of electrons in the core orbitals or boron, NMR chemical shift of boron atom).

Most models, except the tree-based ensemble models, were less performant on uncorrelated features (Table S17 & Table S18). Yet, this is less crucial for our purposes as reliable explanations can be derived from reasonably good predictive models.

Interpretability

	Linear	**LR**	**Bayes. Ridge**	**LASSO**	**SVR**	**Tree**	**RF**	**Grad. Boost.**	**GPR**	**KNN**	**MLP**
All features	16.27	18.09	16.82	16.18	18.16	21.53	14.58	13.39	17.14	20.57	16.1
Uncorrelated features	20.93	21.21	21.01	20.96	21.8	20.69	15.7	14.61	21.22	21.43	20.94

Table S17 : MAEs for selected models evaluated on all database (ONO, NNN, OCO, triarylboranes) with the quantum descriptors reduced or not to uncorrelated features (10-fold cross validation repeated 10 times with different split).

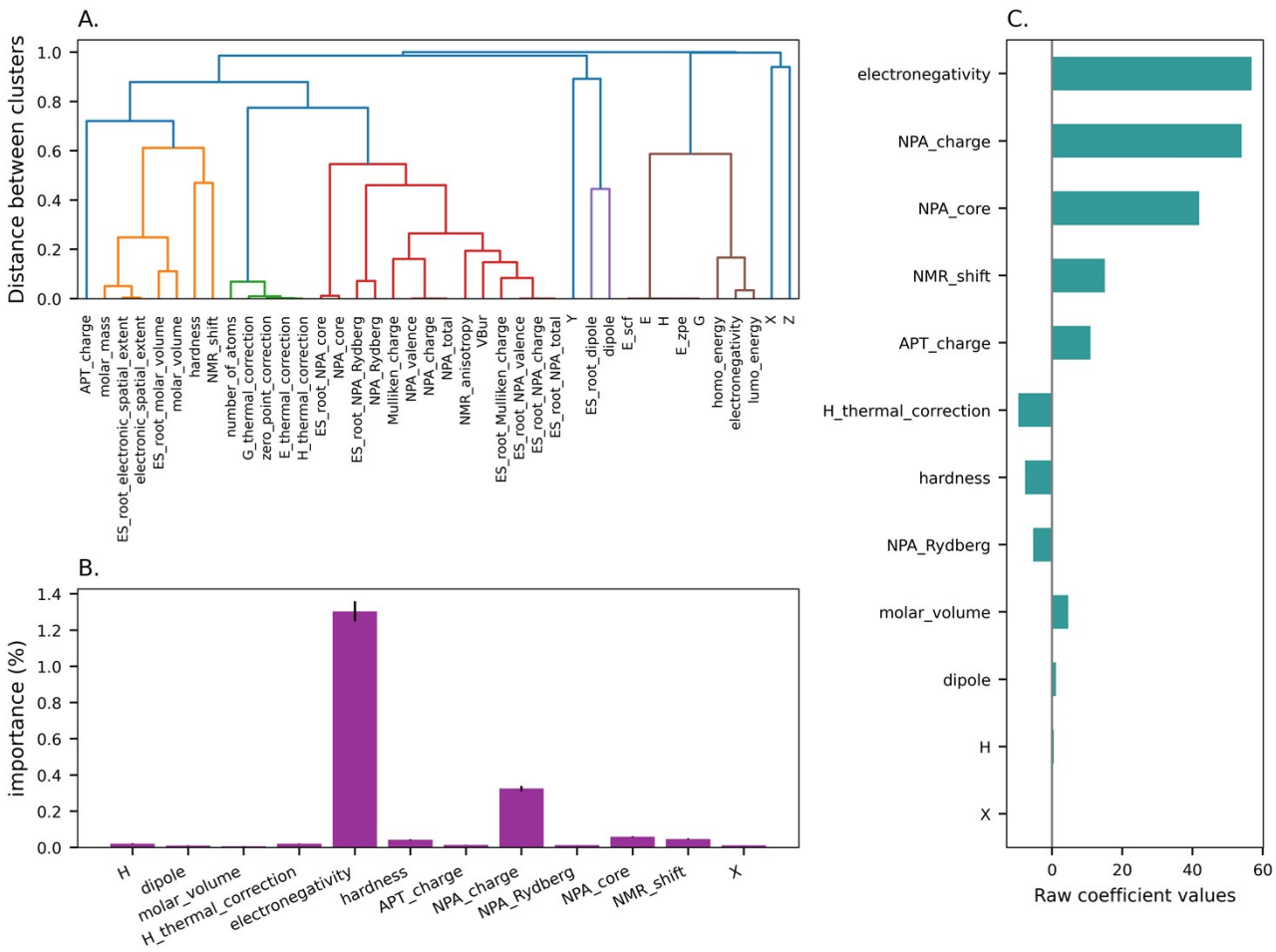


Figure S14 : Interpretability of the quantum features. Models fitted on the whole database (ONO, NNN, OCO, triarylboranes). A. Hierarchical clustering on quantum features. B. Permutation of feature importance for the gradient boosting regressor. C. Linear regression model coefficients.

The high coefficient of core electrons (NPA_core) in the linear regression model is likely an artifact (Figure S14), as it strongly correlates with the molecular scaffolds (Figure S15).

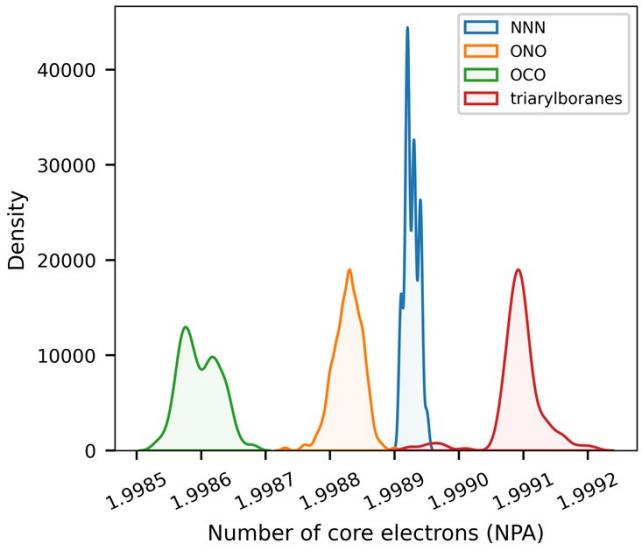


Figure S15 : Distributions of the number of core electrons among the database.

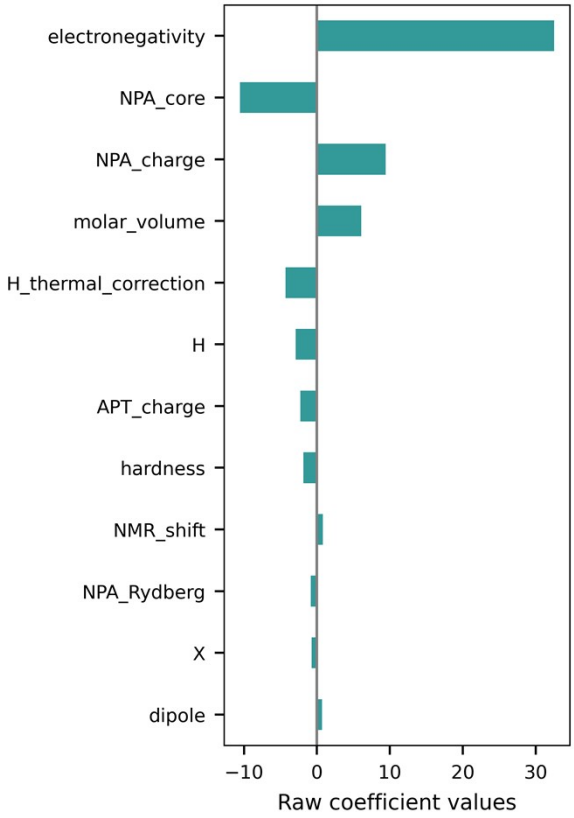


Figure S16 : Coefficients of a linear regression model fitted on uncorrelated features for the ONO molecular scaffold only.

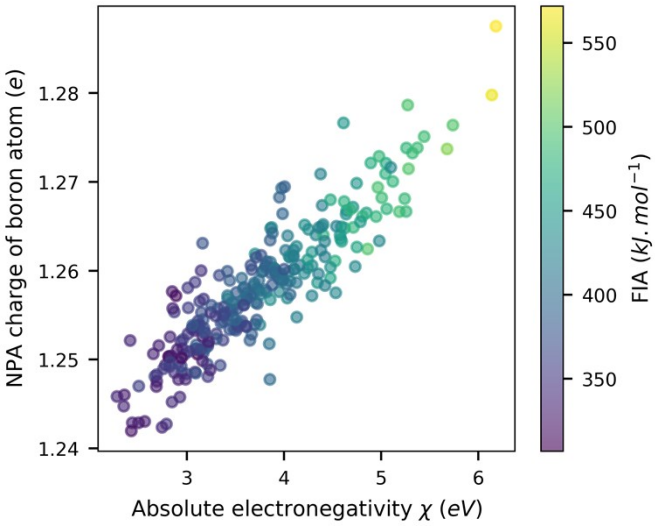


Figure S17 : FIA plot as a function of NPA charge of boron atom and absolute electronegativity of the molecule, for the ONO scaffold. *FIA = 60.0 χ + 8.15 NPA charge + 161*, R^2 = 0.88.

Rationalization of the effect of the nature and position of substituents

Decorrelating features

Hierarchical clustering was performed to decorrelate features (Figure S19 A & C). The threshold was set to 0.55 for the two scaffolds.

Interpretability

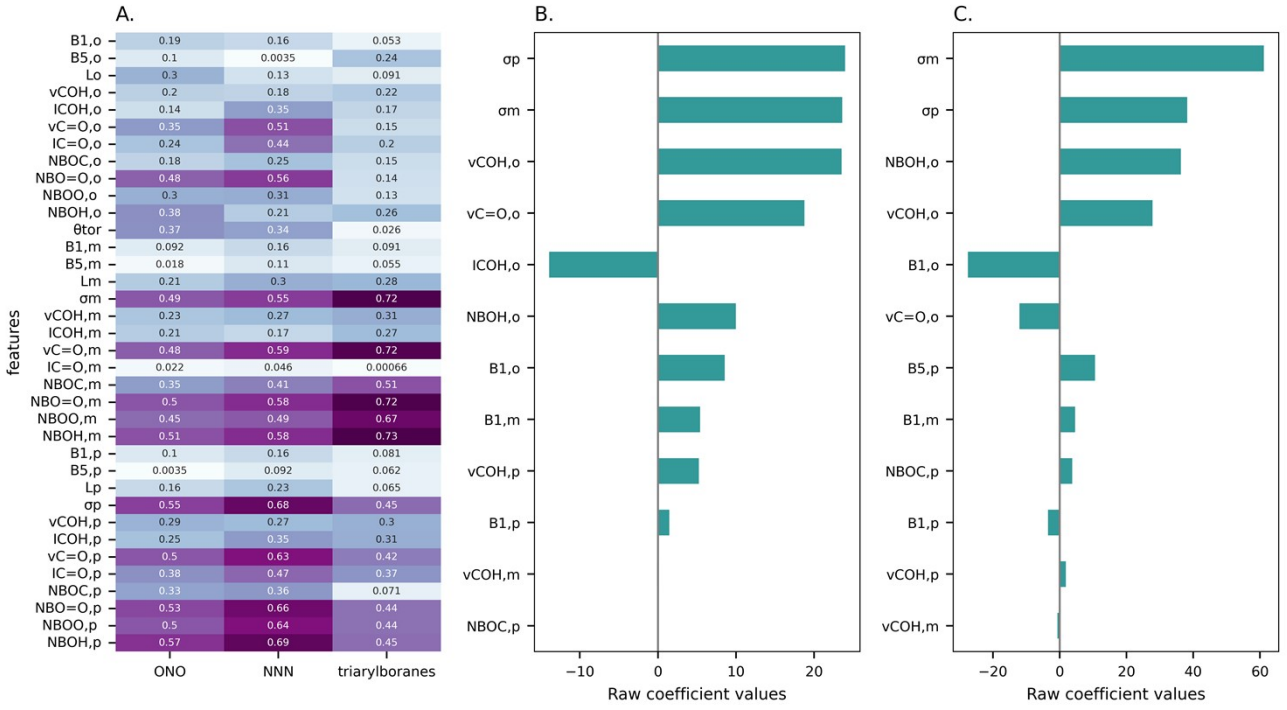


Figure S18 : Analyzing Hammett-extended descriptors. A. Pearson's correlation coefficients of features with FIA for ONO, NNN and triarylboranes molecular scaffolds. B. Coefficients of linear regression model fitted on the ONO dataset using selected uncorrelated features. C. Coefficients of linear regression model fitted on the triarylboranes dataset using selected uncorrelated features.

	Linear	LR	Bayes. Ridge	LASSO	SVR	Tree	RF	Grad. Boost.	GPR	KNN	MLP
ONO	15.27	15.29	15.24	15.23	16.1	14.79	10.56	6.52	12.07	18.94	15.29
Triarylboranes	22.11	22.14	22.1	22.12	25.98	26.87	21.77	16.66	29.21	35.21	22.02

Table S18 : MAEs (kJ.mol^{-1}) of models for ONO and triarylboranes with the uncorrelated Hammett-extended descriptors (10-fold cross-validation repeated 10 times).

We chose gradient boosting regressor model to perform permutation importance as it had the best performance for both scaffolds using the selected uncorrelated features.

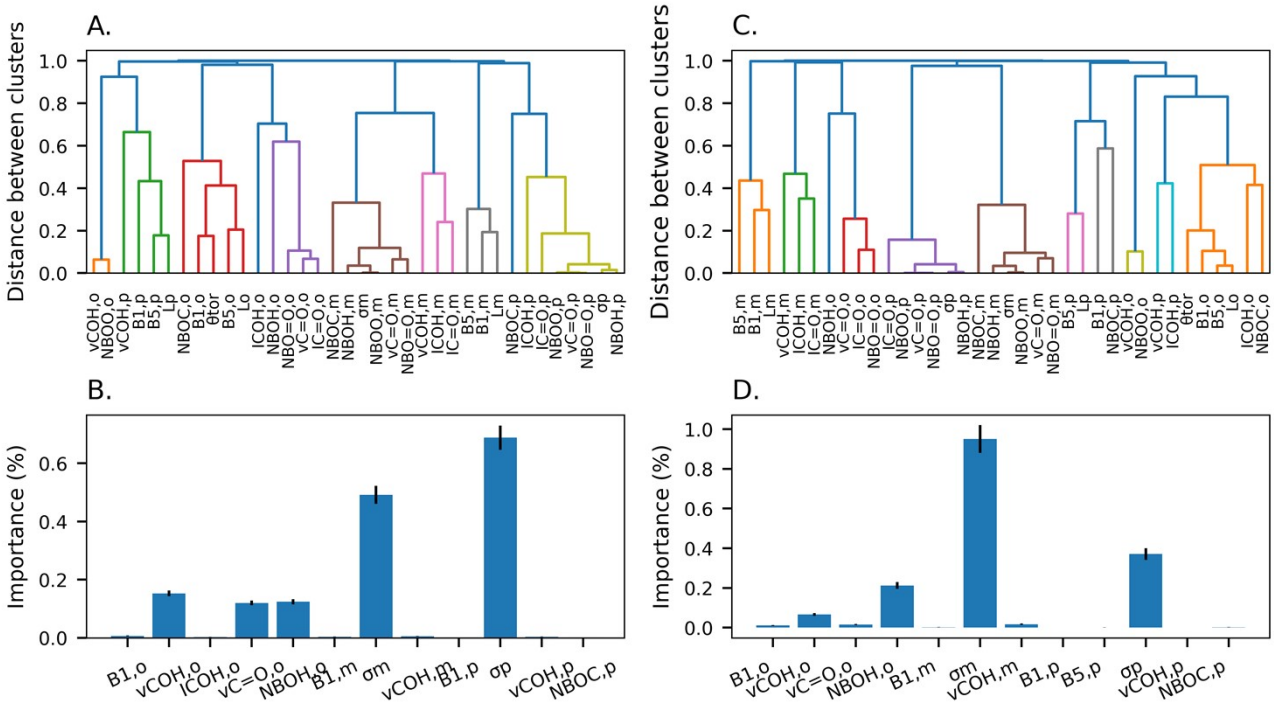


Figure S19 : Permutation importance on Hammett-extended descriptors. A & B.ONO structure. C & D. Triarylboranes.

Comparison between ONO and triarylboranes scaffolds

Analysis of the coefficients of linear model and of the permutation importance algorithm reveals that for the ONO scaffold, σ_p is slightly more important than σ_m, while it is the opposite for triarylboranes, σ_m predominates, followed by σ_p. This indicates that when designing a triarylborane compound, fixing the *meta* group already determines the range of FIA. The final value can be refined by the choice of substituents put at the *ortho* and *para* positions.

Decision trees

The data was not scaled in this part of the analysis to simplify the interpretation of the threshold values at split nodes. Moreover, decision trees are generally insensitive to the scale of the data. A leaf is considered purer (i.e., it contains entries belonging to the same class) when its impurity measure is closer to zero. For this study, models were built using selected Hammett features for simplification. The maximum depth of the trees was limited to 3 for easier interpretation. A criterion on entropy (a measure of the impurity of the leaf) was employed.

ONO scaffold

Classes are: medium LA, good LA, strong LA, super LA and selected features are: $NBO{=}O_o$, $NBO{=}O_m$, $NBO{=}O_p$, and L_o.

The decision tree model was evaluated using 10-fold cross-validation repeated 10 times, achieving an accuracy of 0.73. This value is satisfying to derive reliable interpretations. The model was then fitted on the whole ONO dataset and plotted (Figure S20).

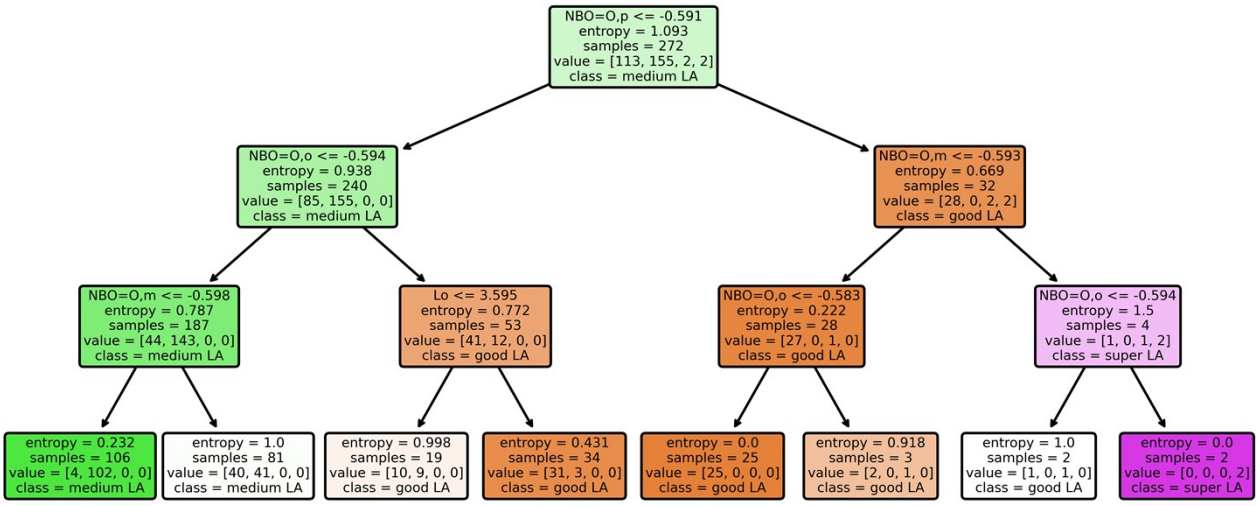


Figure S20 : Scikit-learn[38] tree fitted on the whole dataset for ONO scaffold (0.73 accuracy on 10-fold CV).

Triarylboranes

FIA distribution for triarylboranes being broader, classes are: very weak LA, weak LA, medium LA, good LA, strong LA, super LA. We started using $NBO_{=O}$ partial charges, torsional angle θtor and Lo as the selected features. The cross-validation accuracy score of the resulting tree is 0.48 and $NBO_{=O,\,o}$ and L_o are not used by the model. As the electronic demand of the *ortho* substituent seemed less important, we used only σ_m, σ_p, and replaced Lo by θtor for easier interpretation (accuracy score is then 0.49 on 10-fold CV). The results were less straightforward compared to the ONO dataset (Figure S21). The root node assesses whether the *meta* substituent is electron-donating or not, which partly cleaves the database between good and medium LA (Figure S22). This finding is in agreement with the previous interpretations on Hammett-extended descriptors, namely that the *meta* substituent is the most important to determine FIA value. However, even with an electron-withdrawing group in the *meta* position, medium LA can still be obtained, provided that this group does not exhibit strong mesomeric attraction and is paired with a *para* group that acts as a good donor, such as an amine or tertiobutyl (*t*Bu). Steric effects play a pivotal role in distinguishing between molecules featuring an electron-donating group in the *meta* position and a *para* group with limited electronic attraction (e.g., not NO_2, CN, or CF_3): if the *ortho* group induces sufficient steric hindrance to cause a torsional angle θtor between the carboxylic acid moiety and the ring, the LA tends to be medium or weak; otherwise, it may still be good.

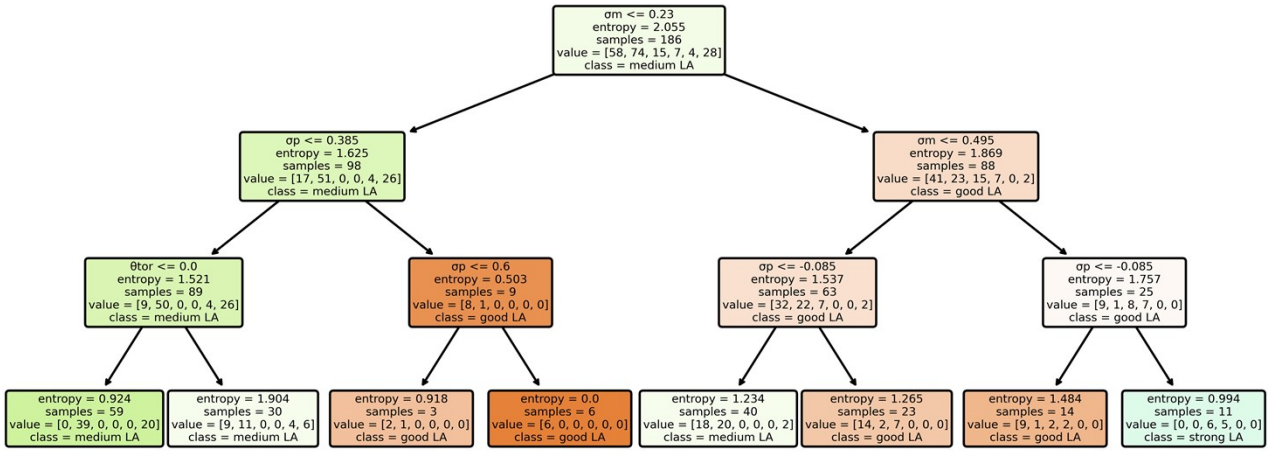


Figure S21 : Triarylboranes Scikit-learn[38] tree.

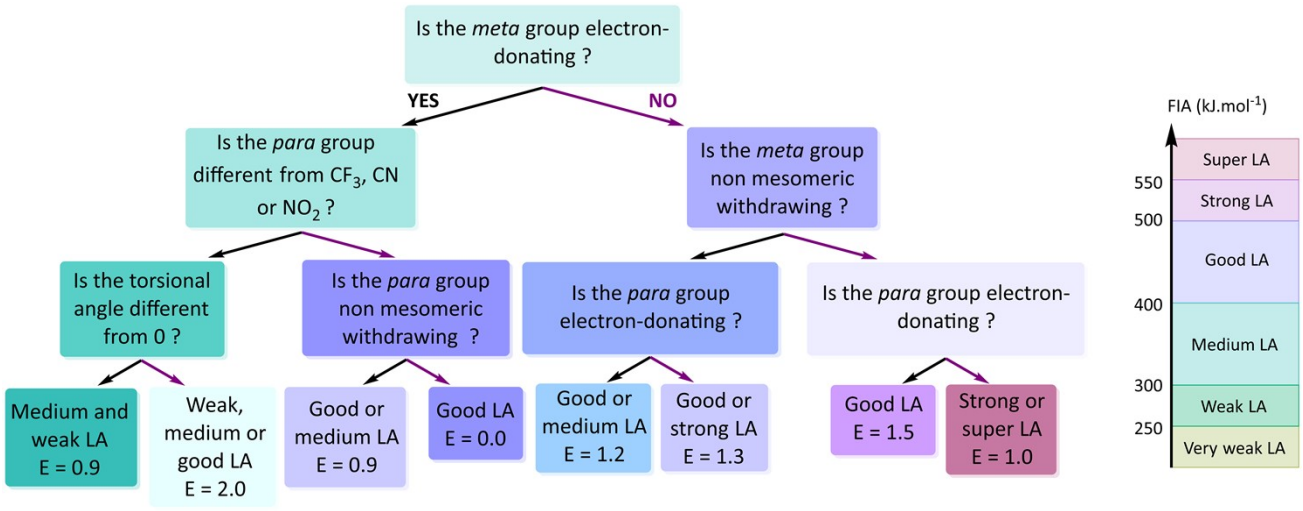


Figure S22 : Tree translation for triarylboranes (0.49 accuracy on 10-fold CV).

References

1 P. Erdmann, J. Leitner, J. Schwarz and L. Greb, *ChemPhysChem*, 2020, **21**, 987–994.
2 L. Greb, *Chem. Eur. J.*, 2018, **24**, 17881–17896.
3 R. G. Parr and R. G. Pearson, *J. Am. Chem. Soc.*, 1983, **105**, 7512–7516.
4 P. Erdmann and L. Greb, *Angewandte Chemie*, DOI:10.1002/ange.202114550.
5 A. R. Jupp, T. C. Johnstone and D. W. Stephan, *Dalton Trans.*, 2018, **47**, 7029–7035.
6 M. A. Beckett, G. C. Strickland, J. R. Holland and K. Sukumar Varma, *Polymer*, 1996, **37**, 4629–4631.
7 R. F. Childs, D. L. Mulholland and A. Nixon, *Can. J. Chem.*, 1982, **60**, 801–808.
8 A. E. Ashley, T. J. Herrington, G. G. Wildgoose, H. Zaher, A. L. Thompson, N. H. Rees, T. Krämer and D. O'Hare, *J. Am. Chem. Soc.*, 2011, **133**, 14727–14740.
9 A. Ben Saida, A. Chardon, A. Osi, N. Tumanov, J. Wouters, A. I. Adjieufack, B. Champagne and G. Berionni, *Angew. Chem. Int. Ed.*, 2019, **58**, 16889–16893.
10 R. C. Bauer, *Synthesis of boracyclobutenes by transmetallation: insertion and demetallation reactions of boracyclobutenes and titanacyclobutene complexes*, Library and Archives Canada = Bibliothèque et Archives Canada, Ottawa, 2010.

11 M. A. Beckett, M. P. Rugen-Hankey, G. C. Strickland and K. S. Varma, *Phosphorus, Sulfur, and Silicon and the Related Elements*, 2001, **169**, 113–116.
12 M. A. Beckett, P. Owen and K. S. Varma, *Journal of Organometallic Chemistry*, 1999, **588**, 107–112.
13 M. Mewald, R. Fröhlich and M. Oestreich, *Chemistry A European J*, 2011, **17**, 9406–9414.
14 M. Ullrich, A. J. Lough and D. W. Stephan, *J. Am. Chem. Soc.*, 2009, **131**, 52–53.
15 M. M. Morgan, A. J. V. Marwitz, W. E. Piers and M. Parvez, *Organometallics*, 2013, **32**, 317–322.
16 T. J. Herrington, A. J. W. Thom, A. J. P. White and A. E. Ashley, *Dalton Trans.*, 2012, **41**, 9019.
17 Z. Lu, Z. Cheng, Z. Chen, L. Weng, Z. H. Li and H. Wang, *Angew Chem Int Ed*, 2011, **50**, 12227–12231.
18 G. J. P. Britovsek, J. Ugolotti and A. J. P. White, *Organometallics*, 2005, **24**, 1685–1691.
19 E. L. Myers, C. P. Butts and V. K. Aggarwal, *Chem. Commun.*, 2006, 4434–4436.
20 A. Adamczyk-Woźniak, M. Jakubczyk, A. Sporzyński and G. Żukowska, *Inorganic Chemistry Communications*, 2011, **14**, 1753–1755.
21 L. A. Körte, J. Schwabedissen, M. Soffner, S. Blomeyer, C. G. Reuter, Y. V. Vishnevskiy, B. Neumann, H. Stammler and N. W. Mitzel, *Angew Chem Int Ed*, 2017, **56**, 8578–8582.
22 C. P. Manankandayalage, D. K. Unruh and C. Krempner, *Dalton Trans.*, 2020, **49**, 4834–4842.
23 L. M. Sigmund, S. S. S., A. Albers, P. Erdmann, R. S. Paton and L. Greb, *Angew Chem Int Ed*, 2024, e202401084.
24 K. O. Christe, D. A. Dixon, D. McLemore, W. W. Wilson, J. A. Sheehy and J. A. Boatz, *Journal of Fluorine Chemistry*, 2000, **101**, 151–153.
25 H. Böhrer, N. Trapp, D. Himmel, M. Schleep and I. Krossing, *Dalton Trans.*, 2015, **44**, 7489–7499.
26 H. Großekappenberg, M. Reißmann, M. Schmidtmann and T. Müller, *Organometallics*, 2015, **34**, 4952–4958.
27 F. Huang, J. Jiang, M. Wen and Z.-X. Wang, *J. Theor. Comput. Chem.*, 2014, **13**, 1350074.
28 J. MacQueen, in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, University of California Press, 1967, vol. 5.1, pp. 281–298.
29 D. J. Woodward, A. R. Bradley and W. P. van Hoorn, *J Chem Inf Model*, 2022, **62**, 4391–4402.
30 D. Weininger, *Journal of Chemical Information and Modeling*, 1988, **28**, 31–36.
31 D. Rogers and M. Hahn, *J. Chem. Inf. Model.*, 2010, **50**, 742–754.
32 RDKit, https://www.rdkit.org/, (accessed 9 November 2022).
33 DeepChem, https://deepchem.io/, (accessed 11 November 2022).
34 A. M. Żurański, J. Y. Wang, B. J. Shields and A. G. Doyle, *React. Chem. Eng.*, 2022, 10.1039.D2RE00030J.
35 Daylight Inc. 4. SMARTS—a language for describing molecular patterns. http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html.
36 XAI_boron_LA https://github.com/jfenogli/XAI_boron_LA.
37 C. B. Santiago, A. Milo and M. S. Sigman, *J. Am. Chem. Soc.*, 2016, **138**, 13424–13430.
38 Scikit-learn https://scikit-learn.org/.