# Quantum Mechanics Enables “Freedom of Design” in Molecular Property Space

**Leonardo Medrano Sandonas[1,*], Johannes Hoja[1,2], Brian G. Ernst[3], Alvaro Vazquez-Mayagoitia[4], Robert A. DiStasio Jr.[3,*], and Alexandre Tkatchenko[1,*]**

[1]Department of Physics and Materials Science, University of Luxembourg, L-1511 Luxembourg, Luxembourg.
[2]Institute of Chemistry, University of Graz, 8010 Graz, Austria.
[3]Department of Chemistry and Chemical Biology, Cornell University, Ithaca, NY 14853, USA.
[4]Computational Science Division, Argonne National Laboratory, Lemont, IL 60439, USA.
[*]Corresponding authors: Leonardo Medrano Sandonas (leonardo.medrano@uni.lu), Robert A. DiStasio Jr. (distasio@cornell.edu), Alexandre Tkatchenko (alexandre.tkatchenko@uni.lu)

## ABSTRACT

Rational design of molecules with targeted properties requires understanding quantum-mechanical (QM) structure-property/property-property relationships (SPR/PPR) across chemical compound space. We analyze these relationships using the QM7-X dataset—which includes multiple QM properties for $\approx 4.2$ M equilibrium and non-equilibrium structures of small (primarily organic) molecules. Instead of providing simple SPR/PPR that strictly follow physicochemical intuition, our analysis uncovers substantial flexibility in molecular property space (MPS) when searching for a single molecule with a desired pair of QM properties or distinct molecules with a targeted set of QM properties. As proof-of-concept, we used Pareto multi-property optimization to search for the most promising (*i.e.*, highly polarizable and electrically stable) molecules for polymeric battery materials; without prior knowledge of this complex manifold of MPS, Pareto front analysis reflected this intrinsic flexibility and identified small directed structural/compositional changes that simultaneously optimize these properties. Our analysis of such extensive QM property data provides compelling evidence for an intrinsic “freedom of design” in MPS, and indicates that rational design of molecules with a diverse array of targeted QM properties is quite feasible.

## 1 Introduction

In recent years, exploration of the vast chemical compound space (CCS) of molecules and materials with data-driven approaches has inspired countless academic and industrial initiatives to seek out the relationships existing between chemical structure and physicochemical properties in this complex high-dimensional space[1–8]. Furthermore, the increasing availability of accurate and reliable molecular property data coupled with the application of sophisticated machine learning (ML) algorithms to explore this data have substantially improved our understanding of quantitative structure-property/property-property relationships (QSPR/QPPR)[9–14]. Such advances have been particularly helpful in the design of novel drugs, antivirals, antibiotics, catalysts, battery materials, and molecules with desired properties[15–21]—processes that have traditionally been driven by chemical intuition or serendipitous discoveries. Despite significant progress in this area, we still lack a comprehensive understanding of the complex relationships that exist (among and) between the structural signatures of molecules (*e.g.*, chemical compositions, conformations) and their physicochemical properties (*e.g.*, energies, forces, HOMO-LUMO gaps, polarizabilities), even in the CCS spanned by small organic molecules only. Unravelling these unknown and complex fundamental SPR/PPR would not only provide us with the tools needed for identifying molecules in high-dimensional molecular property space (MPS), but also the ability to rationally design molecules with a diverse array of targeted physicochemical properties.

To address this challenge, the GDB databases[22–26] have enumerated the molecular graphs comprising large sectors of CCS, enabling us to navigate swaths of CCS that are too vast to be cataloged and studied experimentally. To gain deeper insight into the sector of CCS spanned by small (primarily organic) molecules, several researchers have built upon this work by computing quantum-mechanical (QM) structural and property information corresponding to each molecular graph[27–34]. For instance, the QM7 dataset includes the equilibrium structures of 7,211 small molecules extracted from GDB-13[24] (each containing up to seven heavy/non-hydrogen atoms, including C, N, O, S, and Cl) along with 15 physicochemical properties per molecule at different levels of theory (*i.e.*, ZINDO, SCS, PBE0, GW)[27,28] with variants thereof that computed a number of different molecular properties using (LR-)CCSD.[32] The subsequent QM9 dataset went one step further by generating the structures and 16 (geometric, energetic, electronic, and thermodynamic) properties of 133,885 molecules (each containing up to nine heavy atoms, including C, N, O, and F) from GDB-17[23], all of which were computed at the B3LYP/6-31G(2df,p) level[29]. An even more exhaustive exploration of the CCS of small molecules was accomplished by the ANI-1 dataset[30,31], which consists of more than 20 M equilibrium and non-equilibrium conformations of molecules containing up to eight heavy atoms (including C, N, and O only) from GDB-11[25,26]. More recently, the ANI-1x dataset[33] was also introduced, which contains 20 properties for $\approx 5$ M structures computed using the $\omega$B97-X density functional. Despite all of these foundational efforts to generate a fully QM description of the CCS spanned by small molecules, many challenges exist when translating a series of molecular graphs (which only contain atom connectivity information) to a systematic sampling of CCS which contains an accurate and reliable account of both structural information (*i.e.*, equilibrium and non-equilibrium conformations of constitutional/structural isomers and stereoisomers, including cis-/trans- and conformational isomers) as well as property information (*i.e.*, an extensive and well-converged inventory of QM properties). To address these challenges, the recently published QM7-X dataset[34] provides a systematic, extensive, and tightly converged (PBE0+MBD level of theory) dataset of 42 QM-based physical and chemical properties (including global (molecular), local (atom-in-a-molecule), ground-state, and response properties) for $\approx 4.2$ M equilibrium and non-equilibrium structures of the small molecules in QM7-X, providing what is arguably the most comprehensive account of the CCS spanned by small (primarily organic) molecules to date.

In this work, we performed a comprehensive analysis of the high-dimensional MPS contained in the QM7-X dataset to gain a deeper understanding of the complex SPR/PPR existing in the sector of CCS spanned by small (primarily organic) molecules. In doing so, we found weak correlations existing between most QM properties (*i.e.*, essentially structureless "blobs" in 2D), and in some cases, these relationships went beyond widely accepted chemical and/or physical intuition, *e.g.*, the direct proportionality between molecular size and dispersion energy, inverse proportionality between HOMO-LUMO gap and polarizability, etc. Instead of uncovering simple chemical design rules, our analysis of this extensive QM property database demonstrated that there are very few strict limitations preventing a molecule from exhibiting a desired pair of QM properties. We then investigated even more complex manifolds of MPS and their underlying dependence on molecular structure and chemical composition (*i.e.*, the tunable "knobs" in molecular design), and found multiple cases where two distinct molecules shared multiple QM properties—another indication of the flexibility (or "freedom of design") that one has in the *in silico* search for molecules with a diverse and targeted array of QM properties. Based on these findings, we then employed Pareto front analysis, a powerful multi-property optimization approach, to identify the most promising small organic molecules in CCS (as enumerated by QM7-X) for polymeric battery materials, *i.e.*, molecules with simultaneously large polarizabilities ($\alpha$) and electrical stabilities ($E_{\mathrm{gap}}$). Without any prior knowledge of $\left(\alpha, E_{\mathrm{gap}}\right)$-space, each Pareto front not only reflected this "freedom of design" but also revealed a series of small directed changes to the structure and chemical composition of each Pareto-optimal molecule that simultaneously maximize both of these seemingly contrasting QM properties. We expect that the insight provided in this work will emphasize the critical importance of obtaining high-quality QM property data and contribute to the development of ML-based tools that will considerably improve the sampling, identification, and design of molecular systems for a number of applications, ranging from novel polymeric batteries and organic semiconductors to promising pharmaceuticals and small-molecule protein inhibitors.

## 2 Results

Our comprehensive analysis of the high-dimensional MPS contained in QM7-X (see Figure 1) includes the following four thrusts: (*i*) projecting the 42-dimensional (42D) MPS onto 2D correlation plots for identifying pairwise PPR; (*ii*) characterizing the structural and compositional dependence of global and local properties; (*iii*) exploring more complex manifolds of MPS (*i.e.*, multi-property analysis) by considering the *in silico* design of promising molecules for polymeric battery materials (*i.e.*, highly polarizable and electrically stable molecules with simultaneously large $\langle\alpha\rangle$ and $\langle E_{\mathrm{gap}}\rangle$); (*iv*) finding and analyzing Pareto fronts of molecules with such targeted arrays of properties (*i.e.*, multi-property optimization). For more details about the molecular structures used in these analyses, see *Methods*.

### Pairwise Correlations in Molecular Property Space

As a first step towards understanding the MPS spanned by small (primarily organic) molecules, we analyzed the correlations existing between pairs of properties in QM7-X. To do so, we plotted 2D projections of the 42D QM7-X MPS in Figure 1 for a select subset of 18 properties (including 2 structural, 10 molecular/global, and 6 atom-in-a-molecule/local properties; see Table S1 for more details). In general, Figure 1 shows that the majority of properties do *not* exhibit clear correlations among them; instead, most 2D projections appear as structureless "blobs" indicating very weak (or uncorrelated) PPR, *i.e.*, these properties have a Pearson correlation coefficient $|\rho| < 0.42$ (see *Methods*). In this regard, only four of the 153 (unique pair) projections (*i.e.*, 2.6%; $(C_6, \alpha)$, $(\widetilde{C}_6, \widetilde{\alpha})$, $(\widetilde{C}_6, R_{\mathrm{vdW}})$, $(\widetilde{\alpha}, R_{\mathrm{vdW}})$) display a strong degree of correlation with $|\rho| > 0.92$, while eleven (*i.e.*, 6.9%; $(E_{\mathrm{TB}}, E_{\mathrm{AT}})$, $(E_{\mathrm{TB}}, E_{\mathrm{MBD}})$, $(E_{\mathrm{MBD}}, C_6)$, $(E_{\mathrm{MBD}}, \alpha)$, $(E_{\mathrm{MBD}}, E_{\mathrm{AT}})$, $(E_{\mathrm{AT}}, C_6)$, $(E_{\mathrm{AT}}, \alpha)$, $(\alpha, \alpha_{\mathrm{xx}})$, $(C_6, \alpha_{\mathrm{xx}})$, $(E_{\mathrm{GAP}}, E_{\mathrm{LUMO}})$, and $(E_{\mathrm{GAP}}, E_{\mathrm{HOMO}})$) exhibit a moderate degree of correlation with $0.42 < |\rho| < 0.92$ in which the dispersion in the data is considerably less than a typical "blob". Here, we note in passing that these pairwise correlations also hold when considering just the 41,537 equilibrium structures in QM7-X (see Figure S1 for select examples).

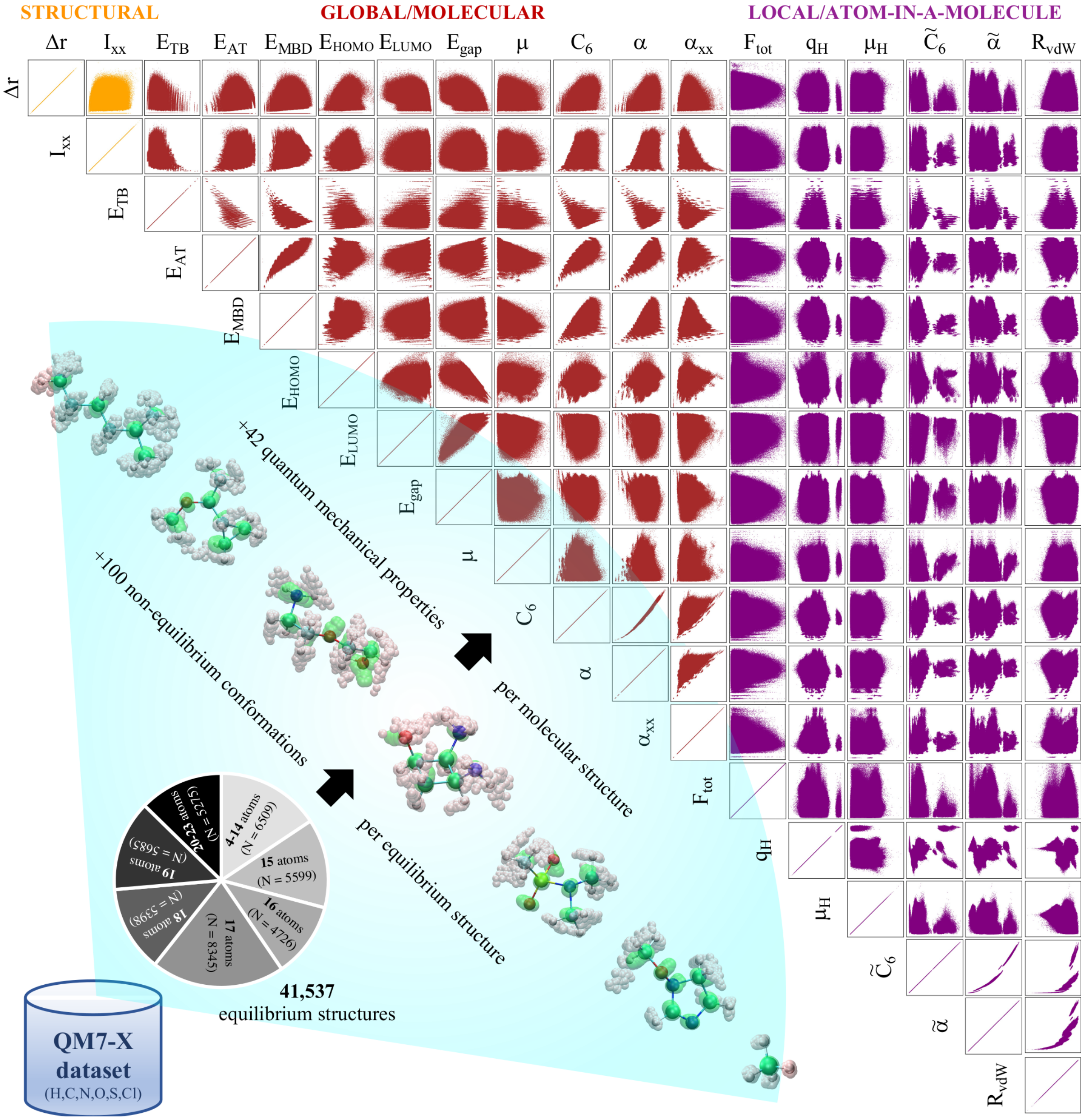

**Figure 1. Pairwise correlations in the QM7-X molecular property space.** The QM7-X dataset[34] includes ≈ 4.2 M (equilibrium and non-equilibrium) molecular structures containing up to seven heavy (C, N, O, S, Cl) atoms, as well as an extensive set of 42 physicochemical properties (per molecular structure) computed using high-level QM calculations (see *Methods* for more details). Select 2D projections of the 42D QM7-X molecular property space (MPS) are depicted for a series of structural (orange), global/molecular (brown), and local/atom-in-a-molecule (violet) properties (see Table S1 for a detailed description of each symbol). Since a vast majority of these correlation plots are structureless "blobs" (*i.e.*, very weak or uncorrelated property-property relationships (PPR)), small primarily organic molecules have the flexibility to exhibit nearly any pair of QM properties.

For instance, consider the 2D projection between $E_{AT}$ and the MBD dispersion energy[35–38] ($E_{MBD}$) in Figure 1; in this case, $E_{AT}$ tends to increase with $E_{MBD}$, in agreement with the expectation that atomization and dispersion energies are extensive properties that increase with molecular size. In this regard, we also observed a moderate correlation between $E_{AT}$ and $\alpha$; although $\alpha$ is non-additive, this quantity does tend to increase with molecular volume[39,40]. The 2D projection between the molecular (isotropic) $C_6$ coefficient and $\alpha$ also indicates a strong correlation; here, the observed quadratic form is rationalized by the Casimir-Polder integral[41], in which the $C_6$ coefficient describing the van der Waals (vdW) interaction between molecules

$A$ and $B$ is given by:

$$C_6 = \frac{3}{\pi}\int_0^\infty d\omega\, \widehat{\alpha}_A(\mathrm{i}\omega)\widehat{\alpha}_B(\mathrm{i}\omega) \approx \frac{3}{2}\left[\frac{\eta_A\eta_B}{\eta_A+\eta_B}\right]\alpha_A\alpha_B, \tag{1}$$

in which $\widehat{\alpha}_{A/B}(\mathrm{i}\omega)$ is the frequency-dependent polarizability of molecule $A/B$ evaluated in the imaginary frequency domain. Substituting the leading-order Padé[42,43] (or quantum harmonic oscillator[44,45]) approximation for $\widehat{\alpha}_{A/B}(\mathrm{i}\omega)$ (*i.e.*, $\widehat{\alpha}_{A/B}(\mathrm{i}\omega) = \alpha_{A/B}/[1-(\omega/\eta_{A/B})^2]$ into this expression yields the well-known London formula in which $C_6 \propto \alpha^2$ (and $\eta_{A/B}$ is the characteristic excitation frequency).

However, seemingly expected correlations (*via* widely accepted chemical and/or physical intuition) were not necessarily observed between global/molecular properties. For example, consider the well-known sum-over-states expression for $\alpha$ from perturbation theory[46,47]:

$$\alpha = 2\sum_{k\neq 0}\frac{|\langle 0|\mu|k\rangle|^2}{E_k-E_0} \approx \frac{|\langle \textsc{homo}|\mu|\textsc{lumo}\rangle|^2}{E_{\mathrm{gap}}}, \tag{2}$$

in which $\langle 0|$ and $|k\rangle$ are the ground (excited) state electronic wavefunctions, $E_0$ ($E_k$) are the corresponding energies, and $\langle 0|\mu|k\rangle$ is the transition dipole moment matrix element. When interpreted using a mean-field one-electron theory (*e.g.*, Hartree-Fock or Kohn-Sham density functional theory), the most significant contribution in Eq. (2) is often the HOMO-LUMO transition. Hence, this sum-over-states expression is commonly approximated (to leading order) by including this term only[48]; within this approximation, $\alpha \propto \frac{1}{E_{\mathrm{gap}}}$, suggesting an inverse proportionality between these properties. While this inverse proportionality can often be observed for a set of homologous molecules (*i.e.*, polyenes[49] and s-*trans* alkenes[50] with increasing length), this relationship does not hold when analyzing the more diverse molecules in QM7-X. Similarly, components of the polarizability tensor (*e.g.*, $\alpha_{\mathrm{xx}}$) appear to be essentially uncorrelated with $C_6$, and do not follow the London formula in Eq. (1) (which corresponds to the scalar/isotropic form of $\alpha$). However, such a lack of correlation between such fundamental molecular properties is by no means uninteresting, and can provide a degree of flexibility that can be exploited in the search for molecules with specific properties, *i.e.*, molecules with preferred polarization directions/orientations to form different molecular crystal polymorphs.

Unlike the global/molecular properties which form single connected "blobs," 2D projections between local/atom-in-a-molecule properties often exhibit distinct clusters, *e.g.*, those involving the Hirshfeld charge ($q_{\mathrm{H}}$), atomic $C_6$ coefficient ($\widetilde{C}_6$), and isotropic atomic polarizability ($\widetilde{\alpha}$) depicted in Figure 1. Such clusters are most visible when analyzing 2D projections between two local properties, and are related to the different atomic environments present in the molecules in QM7-X. For example, the projections involving $q_{\mathrm{H}}$ show the largest number of local atomic environments, and represent the different charge distributions existing in the diverse QM7-X dataset. Local response properties such as $\widetilde{C}_6$, $\widetilde{\alpha}$, and $R_{\mathrm{vdW}}$ (vdW radius) also account for local atomic environments and tend to be strongly correlated. For instance, one can observe multiple quadratic-type functions in the $\left(\widetilde{C}_6, \widetilde{\alpha}\right)$-space, which can be rationalized by the Casimir-Polder relationship applied to each chemical environment (see Eq. (1)). In the same breath, we also find a high degree of correlation between $\widetilde{\alpha}$ and $R_{\mathrm{vdW}}$—a fundamental relationship that has been the topic of discussion in the recent literature[51,52].

With only a handful of exceptions, this analysis of pairwise PPR does not yield simple chemical design rules in the QM7-X sector of CCS spanned by small (primarily organic) molecules. While one might initially view this as a challenge for rational molecule design, this analysis shows that there are very few limitations preventing a molecule from simultaneously exhibiting any desired pair of QM properties. This "freedom of design" hypothesis, which has profound implications in the rational design of molecules with targeted and diverse properties, will be analyzed in more details and confirmed throughout the remainder of this work.

## Structural and Compositional Dependence of Molecular Property Space

The complex set of pairwise PPR found above suggests a certain degree of flexibility in the design of small molecules with a pre-defined set of properties. However, the dependence of the QM7-X MPS on molecular structure and chemical composition—the tunable "knobs" in molecular design—still requires investigation. To do so, we now consider the thermally-averaged ($\langle E_{\mathrm{MBD}}\rangle, \langle E_{\mathrm{AT}}\rangle$)-space as an illustrative probe of this MPS since $\langle E_{\mathrm{MBD}}\rangle$ and $\langle E_{\mathrm{AT}}\rangle$ strongly depend on molecular structure and chemical composition (see *Methods*). Figure 2(a) plots $\langle E_{\mathrm{MBD}}\rangle$ versus $\langle E_{\mathrm{AT}}\rangle$ with each data point colored according to $\langle D_{\max}\rangle$, the *maximum* pairwise distance between heavy/non-hydrogen atoms in a molecular structure. The range of $\langle E_{\mathrm{MBD}}\rangle$ and $\langle E_{\mathrm{AT}}\rangle$ values (0.02−0.48 eV and 19.3−103.3 eV) is quite large, indicating that QM7-X spans a diverse sector of CCS. The molecules with the lowest $\langle E_{\mathrm{MBD}}\rangle$ and $\langle E_{\mathrm{AT}}\rangle$ values are small hydrocarbons such as $CH_4$ ($\sim$ 0.02 eV and $\sim$ 19.3 eV) and $C_2H_2$ ($\sim$ 0.02 eV and $\sim$ 19.9 eV), while the largest values correspond to $C_7H_{16}$ isomers/conformers ($\sim$ 0.48 eV and $\sim$ 103.3 eV); molecules containing second-row atoms (*i.e.*, S and Cl) tend to be characterized by intermediate values.

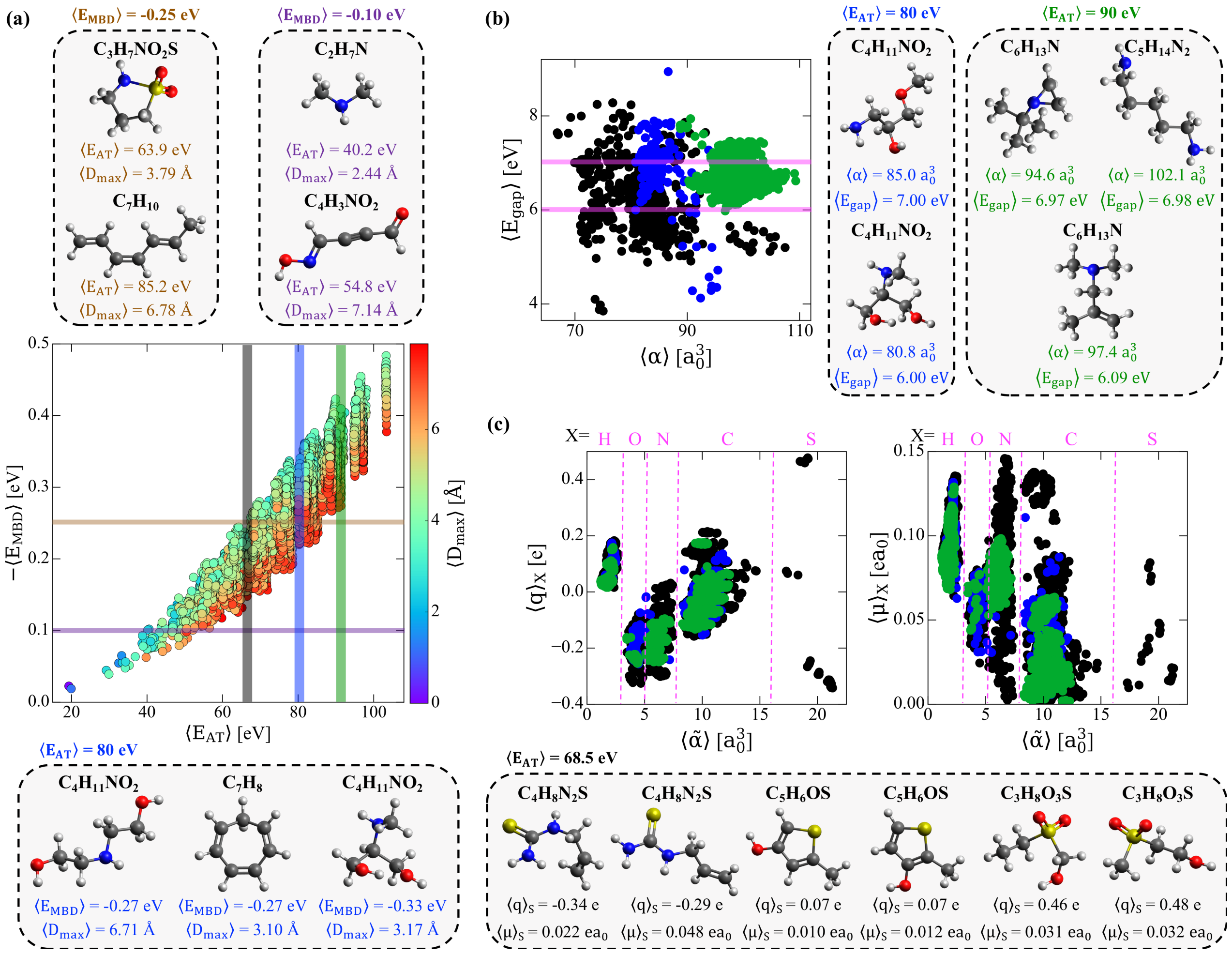

**Figure 2. Structural and compositional dependence of molecular property space: global and local properties. (a)** Correlation plot between the thermally-averaged ($T = 300$ K) MBD dispersion energy ($-\langle E_{\mathrm{MBD}}\rangle$) and atomization energy ($\langle E_{\mathrm{AT}}\rangle$) for the equilibrium structures in QM7-X, with each point colored according to the corresponding thermally-averaged maximum distance between heavy/non-hydrogen atoms ($\langle D_{\mathrm{max}}\rangle$); see *Methods* for more details. Also depicted are select molecules from the $\langle E_{\mathrm{MBD}}\rangle = -0.25 \pm 0.01$ eV and $\langle E_{\mathrm{MBD}}\rangle = -0.10 \pm 0.01$ eV windows (*top inset*) and the $\langle E_{\mathrm{AT}}\rangle = 80 \pm 0.2$ eV window (*bottom inset*). **(b)** 2D projections of the molecules in the highlighted $\langle E_{\mathrm{AT}}\rangle$ windows in panel **(a)** using global/molecular properties ($\langle E_{\mathrm{gap}}\rangle$ and $\langle \alpha\rangle$). Pink lines show the HOMO-LUMO values ($\langle E_{\mathrm{gap}}\rangle = 6.0 \pm 0.05$ eV and $7.0 \pm 0.05$ eV) used during our analysis. Also depicted are select molecules from the $\langle E_{\mathrm{AT}}\rangle = 80 \pm 0.2$ eV and $\langle E_{\mathrm{AT}}\rangle = 90 \pm 0.2$ eV windows (*right inset*). **(c)** 2D projections of the molecules in the highlighted $\langle E_{\mathrm{AT}}\rangle$ windows in panel **(a)** using local/atom-in-a-molecule properties ($\langle q\rangle_{\mathrm{X}}$, $\langle \mu\rangle_{\mathrm{X}}$, and $\langle \widetilde{\alpha}\rangle$). Dashed pink lines are used to delineate the local atomic environments according to the corresponding element (X = H, C, N, O, S). Also depicted are select S-containing molecules from the $\langle E_{\mathrm{AT}}\rangle = 68.5 \pm 0.2$ eV window (*bottom inset*). While global and local properties can be used to distinguish molecules in MPS, our analysis uncovers multiple instances where two distinct molecules (with different structures and compositions) share four or more global properties—compelling evidence for a certain intrinsic flexibility or "freedom of design" in MPS.

Both of these extensive molecular/global properties increase with the number of constituent atoms (independent of chemical composition, see Figure S2), and we again observe a moderate degree of correlation between them (in agreement with that found above when considering all 4.2 M QM7-X structures, *cf.* Figure 1). However, there is considerable dispersion in Figure 2(a), indicating that more diverse ($\langle E_{\mathrm{MBD}}\rangle$, $\langle E_{\mathrm{AT}}\rangle$) combinations are possible, *i.e.*, for a fixed value of one property, there is visible flexibility when choosing the values of the other. From this correlation plot, one can see that this dispersion is fairly well correlated with $\langle D_{\mathrm{max}}\rangle$, a measure of the spatial extent of each molecule. To explore this point further, we characterized the structure and composition of the molecules contained in two fixed $\langle E_{\mathrm{MBD}}\rangle$ windows, *i.e.*, $-0.25 \pm 0.01$ eV and $-0.10 \pm 0.01$ eV, which represent the intermediate-to-low regions of the dispersion energy spectrum (Figure 2(a), top panel). In doing so, one can see that molecules with markedly distinct structures (*i.e.*, compact vs. extended as quantified by $\langle D_{\text{max}} \rangle$) and compositions can have the same $\langle E_{\text{MBD}} \rangle$ but completely different $\langle E_{\text{AT}} \rangle$. This so-called “freedom of design” is clearly illustrated by the $C_3H_7NO_2S$ and $C_7H_{10}$ isomers at the top of Figure 2(a): while both exist in the $\langle E_{\text{MBD}} \rangle = -0.25 \pm 0.01$ eV window, their $\langle E_{\text{AT}} \rangle$ values differ by more than 20 eV. Since $\langle E_{\text{MBD}} \rangle$ is fairly well correlated with the number of atoms in a molecule as well as its volume/spatial extent, $C_7H_{10}$ (an extended molecule with more atoms, $\langle D_{\text{max}} \rangle = 6.78$ Å) and $C_3H_7NO_2S$ (a compact molecule with less atoms, $\langle D_{\text{max}} \rangle = 3.79$ Å) represent a non-trivial compromise between these two effects that results in similar $\langle E_{\text{MBD}} \rangle$ values. In the same breath, the sizeable difference in $\langle E_{\text{AT}} \rangle$ between these molecules can be largely attributed to the larger number of atoms in $C_7H_{10}$ as well as its conjugated/extended $\pi$-system, which further stabilizes this hydrocarbon and increases $\langle E_{\text{AT}} \rangle$. When analyzing the smaller $\langle E_{\text{MBD}} \rangle = -0.10 \pm 0.01$ eV window, we can just as easily find another distinct pair of molecules (again located at the edges of the data dispersion) that exhibit markedly different $\langle E_{\text{AT}} \rangle$ values. Here, we find an $\approx 15$ eV $\langle E_{\text{AT}} \rangle$ difference between $C_2H_7N$ and $C_4H_3NO_2$ (Figure 2(a) top) which can be rationalized by the larger number of heavy atoms and more complex bonding motifs (*e.g.*, C=O, C=N, C≡C) present in $C_4H_3NO_2$.

With such dispersion in $(\langle E_{\text{MBD}} \rangle, \langle E_{\text{AT}} \rangle)$-space, a similar degree of flexibility also exists when holding $\langle E_{\text{AT}} \rangle$ fixed. For instance, analyzing the molecules with $\langle E_{\text{AT}} \rangle = 80 \pm 0.2$ eV uncovered a group of molecules with different structures and/or compositions with the same or different $\langle E_{\text{MBD}} \rangle$ (*e.g.*, the $C_4H_{11}NO_2$ and $C_7H_8$ isomers at the bottom of Figure 2(a)). When comparing the extended (left, $\langle D_{\text{max}} \rangle = 6.71$ Å) and compact (right, $\langle D_{\text{max}} \rangle = 3.17$ Å) $C_4H_{11}NO_2$ isomers, we find that the latter exhibits a more negative $\langle E_{\text{MBD}} \rangle$, consistent with the larger dispersion energy contributions between (closer) non-bonded atoms in compact molecular arrangements. On the contrary, the extended $C_4H_{11}NO_2$ isomer has the same $\langle E_{\text{MBD}} \rangle$ (and $\langle E_{\text{AT}} \rangle$) as the more compact ring-like $C_7H_8$ hydrocarbon ($\langle D_{\text{max}} \rangle = 3.10$ Å)—another illustrative example of the non-trivial compromise between the number of atoms, chemical composition, and volume/spatial extent of a molecule in determining $\langle E_{\text{MBD}} \rangle$. This example also illustrates another aspect of “freedom of design” in MPS, *i.e.*, that two completely distinct molecules can share multiple physicochemical properties (*vide infra*). Interestingly, despite having very similar $\langle D_{\text{max}} \rangle$, the compact $C_4H_{11}NO_2$ isomer has a more negative $\langle E_{\text{MBD}} \rangle$ when compared to the compact but ring-like $C_7H_8$ isomer—a result of more nuanced topological effects (*i.e.*, packed/globular vs. void space) on the dispersion/vdW interactions in molecules[53].

Based on these findings, a natural question arises as to whether or not a similar degree of flexibility exists for other QM properties. To answer this question, we selected three $\langle E_{\text{AT}} \rangle$ windows ($68.5 \pm 0.2$ eV, $80 \pm 0.2$ eV and $90 \pm 0.2$ eV) in Figure 2(a), and analyzed select *global* and *local* PPR among the molecules in these sectors. For global properties, we considered $E_{\text{gap}}$ and $\alpha$, which are important for identifying molecules (with tunable electrical stabilities and polarizabilities) for use in organic electronics and photovoltaic devices. Figure 2(b) depicts the corresponding $\left(\langle E_{\text{gap}} \rangle, \langle \alpha \rangle\right)$ correlation plots (colored according to the $\langle E_{\text{AT}} \rangle$ windows in Figure 2(a)), which appear as structureless "blobs" similar to that found when considering all 4.2 M QM7-X structures (*cf.* Figure 1). Here, we find that the span of $\left(\langle E_{\text{gap}} \rangle, \langle \alpha \rangle\right)$-space is reduced when $\langle E_{\text{AT}} \rangle$ increases and relegated to larger $\langle E_{\text{gap}} \rangle$ and $\langle \alpha \rangle$, implying that molecules with high stabilities to dielectric breakdown *and* enhanced capacities for strong non-covalent interactions can be identified by an initial screen based on $\langle E_{\text{AT}} \rangle$. Since the $\left(\langle E_{\text{gap}} \rangle, \langle \alpha \rangle\right)$-space in Figure 2(b) still contains a large number of molecules, we selected two $\langle E_{\text{gap}} \rangle$ windows ($6.0 \pm 0.05$ eV and $7.0 \pm 0.05$ eV) for further analysis. As a first example, consider the compact $\left(\langle E_{\text{gap}} \rangle = 6.00 \text{ eV}, \langle \alpha \rangle = 80.8\, a_0^3\right)$ and extended $\left(\langle E_{\text{gap}} \rangle = 7.00 \text{ eV}, \langle \alpha \rangle = 85.0\, a_0^3\right)$ $C_4H_{11}NO_2$ isomers in the $\langle E_{\text{AT}} \rangle = 80$ eV window (Figure 2(b)). In this case, the extended isomer has a larger $\langle \alpha \rangle$ despite having a larger $\langle E_{\text{gap}} \rangle$—an illustrative counterexample to the widely used $\alpha \propto \frac{1}{E_{\text{gap}}}$ approximation in Eq. (2). In the same breath, we can just as easily find a pair of isomers that follows this inverse relationship, *i.e.*, the unsaturated $\left(\langle E_{\text{gap}} \rangle = 6.09 \text{ eV}, \langle \alpha \rangle = 97.4\, a_0^3\right)$ and saturated $\left(\langle E_{\text{gap}} \rangle = 6.97 \text{ eV}, \langle \alpha \rangle = 94.6\, a_0^3\right)$ $C_6H_{13}N$ isomers in the $\langle E_{\text{AT}} \rangle = 90$ eV window. Another interesting finding is the $4.7-7.5\, a_0^3$ enhancement in $\langle \alpha \rangle$ when morphing from $C_6H_{13}N$ to $C_5H_{14}N_2$—a clear example of the non-additivity in $\alpha$ (whose role is often underestimated in small molecules) as $\langle \widetilde{\alpha} \rangle_{\text{C}} \approx \langle \widetilde{\alpha} \rangle_{\text{N}} + \langle \widetilde{\alpha} \rangle_{\text{H}}$. From the perspective of Eq. (2), this polarizability enhancement is even more surprising, as $C_5H_{14}N_2$ has a $\langle E_{\text{gap}} \rangle$ that is larger than (or equal to) the $C_6H_{13}N$ isomers. Such an increase in $\langle \alpha \rangle$ is non-trivial and has substantial implications for non-covalent interactions involving these molecules, as $C_6 \propto \alpha^2$ (*cf.* Eq. (1)).

From this analysis, we also found multiple cases where two molecules with markedly different structures and compositions share four (extensive and intensive) global properties, further demonstrating the flexibility one has when designing molecules with an array of targeted properties. As an illustrative example, consider again the saturated $C_6H_{13}N$ and $C_5H_{14}N_2$ isomers in Figure 2(b), which have similar $\langle E_{\text{MBD}} \rangle \approx 0.36 \pm 0.02$ eV and $\langle \mu \rangle \approx 0.26 \pm 0.02$ eÅ (in addition to $\langle E_{\text{AT}} \rangle$ and $\langle E_{\text{gap}} \rangle$). Hence, additional properties (*i.e.*, $\langle \alpha \rangle$ and $\langle D_{\text{max}} \rangle$) are needed to uniquely identify molecules in high-dimensional QM7-X MPS.

In the same breath, local/atom-in-a-molecule properties can also be used to distinguish molecules in MPS. To demonstrate this, we analyzed the molecules in the three $\langle E_{\text{AT}} \rangle$ windows in Figure 2(a) by partitioning them according to $\langle q_{\text{H}} \rangle$, $\langle \mu_{\text{H}} \rangle$, and $\langle \widetilde{\alpha} \rangle$; to enable an atom-specific discussion, the subscript H will be removed from all Hirshfeld quantities (*i.e.*, $\langle q_{\text{H}} \rangle \to \langle q \rangle$, $\langle \mu_{\text{H}} \rangle \to \langle \mu \rangle$), and $\langle q \rangle_{\text{X}}$ ($\langle \mu \rangle_{\text{X}}$) will now refer to the thermally averaged Hirshfeld charge (dipole) on atom X. As depicted in Figure 2(c), we again observe significant clustering in $(\langle q \rangle, \langle \widetilde{\alpha} \rangle)$-space and $(\langle \mu \rangle, \langle \widetilde{\alpha} \rangle)$-space, reflecting the diverse chemical environments in this subset of QM7-X. By delimiting the sectors belonging to each element (X = H, O, N, C, S), we also found

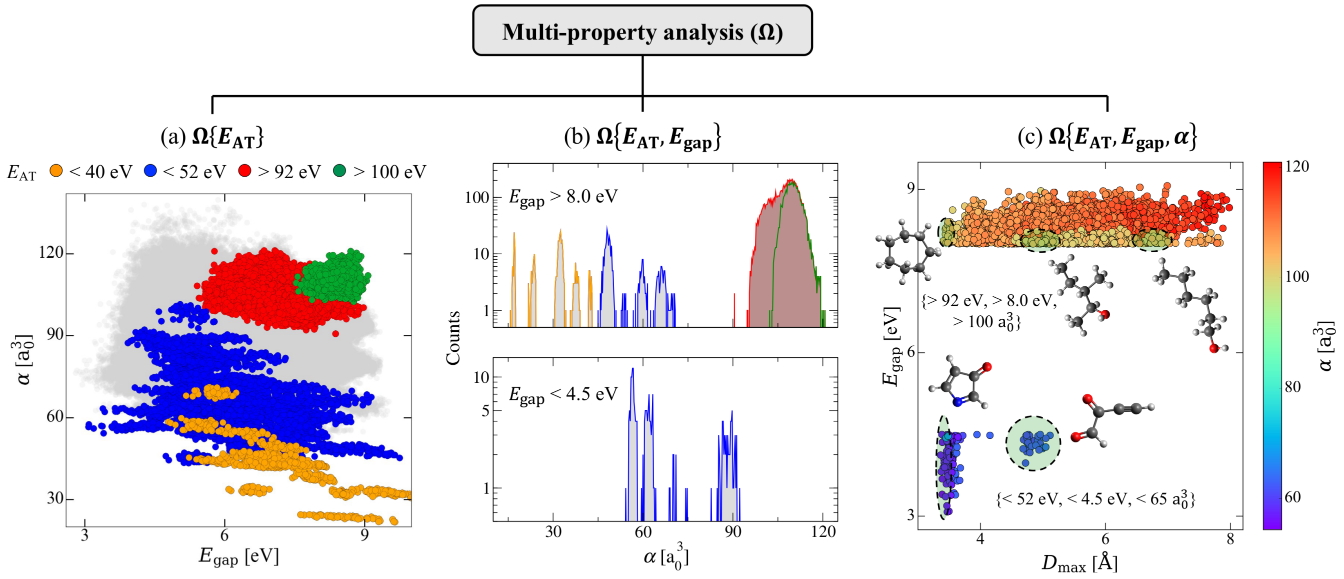

**Figure 3. Multi-property analysis in molecular property space.** Three different multi-property analyses ($\Omega$) of the manifold of MPS defined by $E_{AT}$, $E_{gap}$, and $\alpha$ were performed in which the molecules in QM7-X were progressively partitioned according to an increasing number of properties: **(a)** $\Omega\{E_{AT}\}$ (one property), **(b)** $\Omega\{E_{AT}, E_{gap}\}$ (two properties), and **(c)** $\Omega\{E_{AT}, E_{gap}, \alpha\}$ (three properties). **(a)** Correlation plot between $\alpha$ and $E_{gap}$, with each point colored according to the corresponding $E_{AT}$ range. Light gray points correspond to molecules with $E_{AT} \in [52, 92]$ eV. **(b)** Plots of frequency versus $\alpha$ for molecules with $E_{gap} < 4.5$ eV and $E_{gap} > 8.0$ eV, with each distribution colored according to the $E_{AT}$ ranges in panel **(a)**. **(c)** Correlation plot between $E_{gap}$ and $D_{max}$ for molecules with $\{E_{AT} > 92$ eV, $E_{gap} > 8.0$ eV, $\alpha > 100\ a_0^3\}$ and $\{E_{AT} < 52$ eV, $E_{gap} < 4.5$ eV, $\alpha < 65\ a_0^3\}$, with each point colored according to the corresponding $\alpha$ value. Also depicted are select molecules from each of these sectors of $\left(E_{AT}, E_{gap}, \alpha\right)$-space. From this tiered multi-property analysis, we find further evidence of the "freedom of design" that exists across wide swaths of MPS—there exists a number of molecules with different structures and chemical compositions that share an array of physicochemical properties (*e.g.*, simultaneously large $E_{AT}$ and $E_{gap}$ values).

that H and S display the smallest ($1-3\ a_0^3$) and largest ($16-22\ a_0^3$) $\langle\widetilde{\alpha}\rangle$, respectively. Since the molecules with $\langle E_{AT}\rangle = 68.5$ eV are partitioned into three well-defined clusters in $(\langle q\rangle, \langle\widetilde{\alpha}\rangle)$-space, we focus our discussion on the diverse S-containing molecules in this subregion of QM7-X. Such well-defined clusters reflect the the number of different chemical environments surrounding each S atom; depending on the local charge distribution, this versatile third-row element can act as an electron acceptor ($\langle q\rangle_S < 0$) or donor ($\langle q\rangle_S > 0$), or remain essentially neutral ($\langle q\rangle_S \approx 0$), see Figure 2(c) bottom. In contrast to $\langle\mu\rangle$, $\langle q\rangle$ seems to be a more sensitive probe of the local chemical environment (and charge distribution) surrounding each atom, and may therefore be a useful local property for identifying molecules in high-dimensional QM7-X MPS.

## Multi-Property Analysis: Exploring More Complex Manifolds of Molecular Property Space

Since molecular design often involves the simultaneous optimization of multiple (typically more than two) physicochemical properties, we continue our analysis by exploring more complex manifolds of MPS. As an illustrative example, we consider the *in silico* design of promising molecules for polymeric battery materials (*i.e.*, highly polarizable and electrically stable molecules with simultaneously large $\langle\alpha\rangle$ and $\langle E_{gap}\rangle$)[54,55]. To accomplish this goal, we carried out three different multi-property analyses ($\Omega$) in which the molecules in QM7-X are progressively partitioned according to an increasing number of the following global/molecular properties: $E_{AT}$, $E_{gap}$, and $\alpha$. At the single-property $\Omega\{E_{AT}\}$ level, the QM7-X molecules are partitioned according to pre-defined $E_{AT}$ ranges (see Figure S3 for analogous $\Omega\{E_{gap}\}$ and $\Omega\{\alpha\}$ analyses); after doing so, a number of discernible trends emerge despite the fact that a plot of the total (unpartitioned) $\left(\alpha, E_{gap}\right)$-space has no visible correlation between these properties (see Figure 3(a)). For one, if we just consider molecules with $E_{AT} < 52$ eV (blue points in Figure 3(a)), we find that these molecules exhibit a remarkably wide range of $\alpha$ ($\in [16.5, 101.2]\ a_0^3$) and $E_{gap}$ ($\in [3.0, 11.6]$ eV) values, while the molecules with $E_{AT} > 92$ eV (red points) are more likely to have large $\alpha$ and $E_{gap}$. In both cases, the molecules exhibit considerable flexibility in their $\alpha$ and $E_{gap}$ values, which bodes well for identifying promising polymeric battery material candidates and again illustrates that small organic molecules do not necessarily follow the $\alpha \propto \frac{1}{E_{gap}}$ relationship espoused by Eq. (2).

We continue by partitioning the molecules in QM7-X according to $E_{AT}$ (using the partitions outlined above) and $E_{gap}$

(coarse-grained partitioning into small and large $E_{\mathrm{gap}}$), *i.e.*, multi-property analysis at the $\Omega\{E_{\mathrm{AT}}, E_{\mathrm{gap}}\}$ level. As depicted in Figure 3(b), we find that molecules with $E_{\mathrm{gap}} < 4.5$ eV (small gap) lacked diversity in both structure and composition, and were constrained to intermediate $\alpha$ ($\in [54.3, 92.3]\ a_0^3$). On the other hand, the molecules with $E_{\mathrm{gap}} > 8.0$ eV (large gap) span a significantly wider range of $\alpha$ ($\in [15.8, 120.8]\ a_0^3$) and their structure/composition is noticeably more diverse.

At the $\Omega\{E_{\mathrm{AT}}, E_{\mathrm{gap}}, \alpha\}$ level, we further partition the molecules in Figure 3(b) according to $\alpha$. In the region of $\left(E_{\mathrm{AT}}, E_{\mathrm{gap}}, \alpha\right)$-space defined by small atomization energies ($E_{\mathrm{AT}} < 52$ eV), small HOMO-LUMO gaps ($E_{\mathrm{gap}} < 4.5$ eV), and small polarizabilities ($\alpha < 65\ a_0^3$), we find relatively few molecules and a general lack of diversity in structure, composition, and size (Figure 3(c)). More specifically, we found only 146 molecular structures in this manifold (originating from three different chemical compositions) with a rather limited range of molecular sizes (*i.e.*, 3.4 Å $< D_{\max} <$ 5.1 Å)—in this case, simultaneously restricting the range of these three properties significantly (and not unexpectedly) constrains the molecular design space. On the other hand, the molecules in QM7-X with large atomization energies ($E_{\mathrm{AT}} > 92$ eV), high electrical stabilities ($E_{\mathrm{gap}} > 8.0$ eV) and high polarizabilities ($\alpha > 100\ a_0^3$) are markedly more diverse (7,365 structures, 3.5 Å $< D_{\max} <$ 8.0 Å, see Figure 3(c)). Furthermore, we expect that the molecules in this sector of $\left(E_{\mathrm{AT}}, E_{\mathrm{gap}}, \alpha\right)$-space would be even more diverse (and potentially more promising polymeric battery material candidates) if CCS was probed with an even larger molecular database. From this tiered multi-property analysis, we find that it is feasible to design molecules with completely different structures and compositions that share an array of different physicochemical properties—yet another manifestation of the freedom of design that exists across wide swaths of MPS.

### Multi-Property Optimization: Finding Optimal Pareto Fronts in Molecular Property Space

When optimizing multiple objective functions among a large candidate pool, Pareto fronts (or frontiers) represent the so-called Pareto-optimal solutions for which no single objective function can be improved without degrading the others. Pareto fronts have been used in a number of fields (*e.g.*, economics, medicine, materials science, chemical engineering)[56–60] and have given rise to evolutionary multi-objective optimization[61,62]. In this work, we extend our analysis in the previous section by using this approach to *identify* the most promising small organic molecules in CCS (as enumerated by the QM7-X database) to form polymeric battery materials[54,55], *i.e.*, the Pareto front of molecules in QM7-X which simultaneously have the largest $\langle\alpha\rangle$ and $\langle E_{\mathrm{gap}}\rangle$ values (see *Methods*). Here, we note that this approach is general and could be used to search for molecules with any number/combination of properties (*e.g.*, promising small-molecule protein inhibitors with large $\langle\alpha\rangle$ and reduced $\langle\mu\rangle$ values).

In Figure 4, Pareto fronts with simultaneously optimal $\langle\alpha\rangle$ and $\langle E_{\mathrm{gap}}\rangle$ values are provided for three different $\langle E_{\mathrm{AT}}\rangle$ ranges (enabling us to explore different chemical compositions in each front). Overall, these fronts generally follow the inverse $\alpha \propto \frac{1}{E_{\mathrm{gap}}}$ relationship in Eq. (2); however, there are exceptions and unexpected structures that appear along each front, reflecting the freedom one has when designing molecules with an array of targeted properties. The $\textcircled{A}\rightarrow\textcircled{B}$ front corresponds to molecules with $\langle E_{\mathrm{AT}}\rangle \in [40, 50)$ eV and contains 11 diverse structures with varied compositions, starting with (C,N,S)-based molecules with large $\langle\alpha\rangle$ and (relatively) small $\langle E_{\mathrm{gap}}\rangle$, and ending with simpler and more compact molecules with substantially lower $\langle\alpha\rangle$ and very large $\langle E_{\mathrm{gap}}\rangle$. The first three structures are constitutional isomers of $C_4H_2N_2S$ with a terminal alkyne (ethynyl group) directly adjacent to an aromatic thiadiazole ring. Such conjugation facilitates charge delocalization and $\pi$-electron mobility across each molecule; as such, these isomers have large (but similar) $\langle\alpha\rangle \approx 86.0\ a_0^3$ and (relatively) small gaps. However, the $\langle E_{\mathrm{gap}}\rangle$ are more sensitive to the relative positions of the heteroatoms in the thiadiazole ring and can differ by 0.6 eV. Continuing along the $\textcircled{A}\rightarrow\textcircled{B}$ front, we find a linear yet highly conjugated molecule (penta-2,4-diynenitrile, $C_5HN$), with a structure and composition completely different from the $C_4H_2N_2S$ isomers. Despite such differences, $\langle\alpha\rangle$ and $\left\langle E_{\mathrm{gap}}\right\rangle$ for this molecule are very similar to the Pareto-adjacent $C_4H_2N_2S$ isomers, which again illustrates the flexibility inherent to MPS. Ethynyl sulfone ($C_4H_2O_2S$) is the next molecule, which contains two terminal alkynes connected *via* a central sulfonyl ($SO_2$) moiety in a kinked arrangement—a large change in both structure and composition when compared to $C_5HN$. This non-linear molecular geometry significantly reduces $\pi$-electron mobility and charge delocalization, and results in a rather large ($>$ 1 eV) $\left\langle E_{\mathrm{gap}}\right\rangle$ change. In the same breath, this molecule has a very similar $\langle\alpha\rangle$ value to $C_5HN$—an interesting example of how the concept of "freedom of design" naturally emerges from Pareto front analysis. Even more interesting is how Pareto front analysis can be used to facilitate rational *in silico* design of molecules with targeted properties. To see this, consider the small directed changes needed to arrive at the next three Pareto-optimal molecules in the $\textcircled{A}\rightarrow\textcircled{B}$ front. Since the polarizability of a N atom is smaller than that of a C−H group, replacing an ethynyl (C≡C−H) group in $C_4H_2O_2S$ with a nitrile (C≡N) group (*i.e.*, $C_4H_2O_2S \rightarrow C_3HNO_2S$) can be used to design a molecule with a lower $\langle\alpha\rangle$. Since the central sulfonyl group provides an effective conduit for charge delocalization in $C_3HNO_2S$, replacing $SO_2$ with a more insulating methylene ($CH_2$) group (*i.e.*, $C_4H_2O_2S \rightarrow C_5H_4$) decreases $\langle\alpha\rangle$ by ≈ 12% and increases $\left\langle E_{\mathrm{gap}}\right\rangle$ by ≈ 1 eV. Finally, making *both* replacements (*i.e.*, $C_4H_2O_2S \rightarrow C_4H_3N$) leads to further (and rather predictable) changes in both $\langle\alpha\rangle$ and $\left\langle E_{\mathrm{gap}}\right\rangle$. These small but rational changes to the structure and composition of these molecules are well-aligned with "chemical intuition" and emerged from this analysis without prior knowledge of $\left(\alpha, E_{\mathrm{gap}}\right)$-space; as such, we would argue that Pareto front analysis has tremendous potential in the

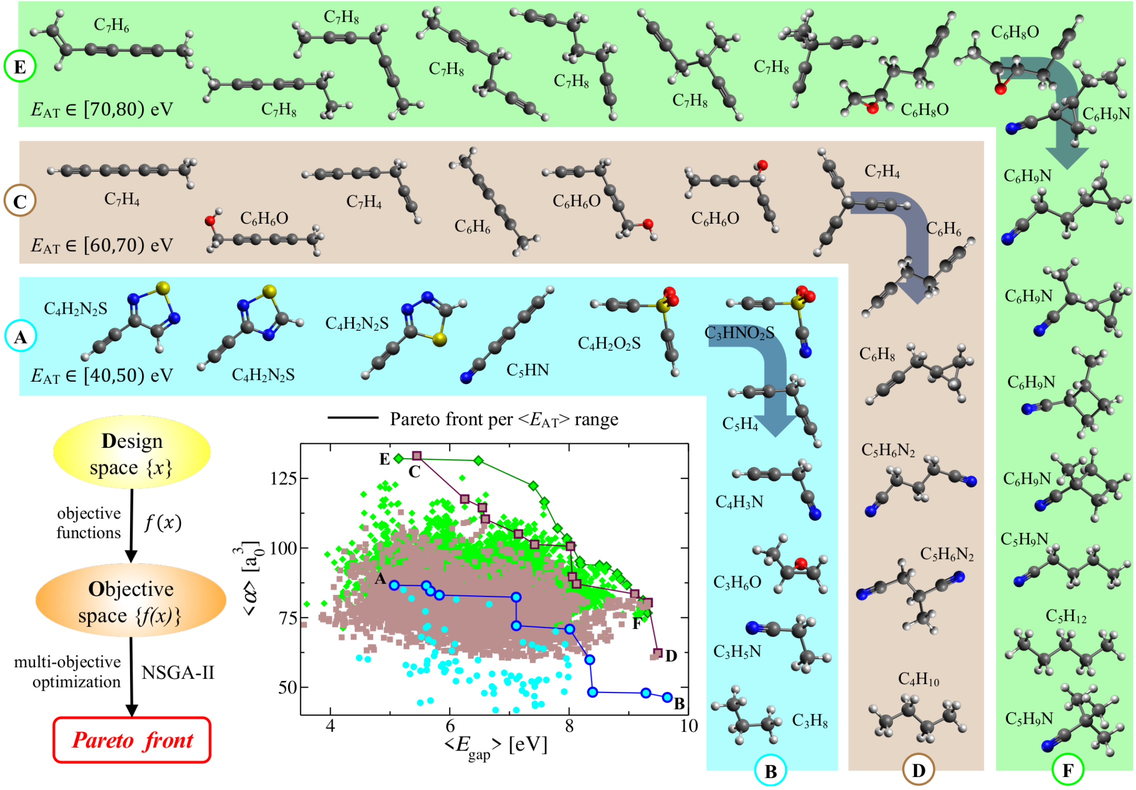

**Figure 4. Multi-property optimization: Pareto front analysis in molecular property space.** Pareto (multi-property) optimization in the manifold of MPS defined by $E_{\text{AT}}$, $E_{\text{gap}}$, and $\alpha$ was performed to identify molecules with simultaneously large $\alpha$ and $E_{\text{gap}}$ values (*i.e.*, an illustrative example for the *in silico* design of promising candidate molecules for polymeric battery materials). Depicted is a correlation plot between $\langle\alpha\rangle$ and $\langle E_{\text{gap}}\rangle$, with each point colored according to the corresponding $E_{\text{AT}}$ range (cyan: $\langle E_{\text{AT}}\rangle \in [40,50)$ eV; tan: $\langle E_{\text{AT}}\rangle \in [60,70)$ eV; light green: $\langle E_{\text{AT}}\rangle \in [70,80)$ eV). The optimal Pareto fronts corresponding to each of these $\langle E_{\text{AT}}\rangle$ windows are provided as highlighted points (connected *via* solid lines) in this correlation plot; see *Methods* for a more detailed description of the Pareto optimization procedure. Also depicted are the QM7-X molecules located on each of these Pareto fronts (and corresponding to the highlighted points in the correlation plot), which reflect the intrinsic flexibility in MPS as well as the small directed structural/compositional changes that are needed in the rational design of molecules with an array of targeted QM properties.

field of *in silico* molecular design. The last segment of this front is somewhat unsurprising and comprised of three simpler and more compact molecules ($C_3H_6O \rightarrow C_3H_5N \rightarrow C_3H_8$), all of which exhibit small (but similar) $\langle\alpha\rangle$ and relatively large $\langle E_{\text{gap}}\rangle$.

To search for even larger candidate molecules, we performed a similar analysis on the QM7-X molecules with $\langle E_{\text{AT}}\rangle \in [60,70)$ eV and $\langle E_{\text{AT}}\rangle \in [70,80)$ eV. In doing so, we again find numerous examples illustrating the flexibility woven into MPS as well as the fact that Pareto front analysis is a powerful (and largely underutilized) tool for *in silico* molecular design. In the $\langle E_{\text{AT}}\rangle \in [60,70)$ eV sector, the Ⓒ→Ⓓ front is essentially a straight line (with 12 molecules) mirroring the inverse relationship between $\langle\alpha\rangle$ and $\langle E_{\text{gap}}\rangle$ with a few exceptions. For example, consider the sixth, seventh, and eighth molecules in the Ⓒ→Ⓓ front in Figure 4 (*i.e.*, $C_6H_6O$, $C_7H_4$, and $C_6H_6$). Here, we find a sharp increase in $\langle E_{\text{gap}}\rangle$ accompanied by almost no change in $\langle\alpha\rangle$ as we move from $C_6H_6O$ (a kinked molecule with two alkynes connected by a central alcohol moiety) to $C_7H_4$ (a propeller-like molecule with three terminal alkynes connected by a central aliphatic CH group). This can be rationalized by the additional non-conjugated triple bond in $C_6H_6O$, which localizes the $\pi$ electrons and increases $\langle E_{\text{gap}}\rangle$. At this point, we also arrive at the front edge, *i.e.*, among all QM7-X molecules with $\langle E_{\text{AT}}\rangle \in [60,70)$ eV, this propeller-like $C_7H_4$ isomer has simultaneously optimal $\langle\alpha\rangle$ and $\langle E_{\text{gap}}\rangle$. Next, we observe a sharp decrease in $\langle\alpha\rangle$ accompanied by almost no change in $\langle E_{\text{gap}}\rangle$ as we move from $C_7H_4$ to $C_6H_6$ (a more extended but staggered molecule with two terminal alkynes connected by a central insulating ethylene ($-CH_2-CH_2^-$) group). This transition maintains the locality of the $\pi$ electrons and is accompanied by the loss of a C atom and the gain of two H atoms; since $\tilde{\alpha}_C > 2\tilde{\alpha}_H$, this will tend to decrease $\langle\alpha\rangle$. In the $\langle E_{\text{AT}}\rangle \in [70,80)$ eV sector, the Ⓔ→Ⓕ front ($N = 17$) is more parabolic in shape. This front is populated by large stretches of structural/constitutional isomers, peppered with local functional group changes, all of which reflect the key aspects of rational molecular design discussed above. For brevity, we leave a more detailed analysis of this front to the interested reader.

## 3 Discussion

The recently developed QM7-X dataset—which includes 42 physicochemical properties obtained *via* high-level QM calculations for approximately 4.2 M (equilibrium and non-equilibrium) molecular structures containing up to seven heavy atoms—allows us to study the sector of CCS spanned by small (primarily organic) molecules. By performing 2D projections of the high-dimensional MPS described by QM7-X, we gained tremendous insight into the complex QM-based SPR/PPR existing in this region of CCS. In general, we found weak correlations among the majority of QM properties considered herein (*i.e.*, essentially structureless "blobs" in 2D), although we did find more nuanced relationships that go beyond chemical and/or physical intuition in some cases. For instance, we did not observe the widely accepted inverse relationship between molecular polarizability and HOMO-LUMO gap ($\alpha \propto \frac{1}{E_{\text{gap}}}$), *i.e.*, the leading-order approximation for $\alpha$ in Eq. (2). Instead of uncovering simple chemical design rules, our analysis of this extensive QM property database demonstrated that there are generally no hard-and-fast limitations preventing a molecule from exhibiting a desired pair of QM properties. In other words, there exists a certain "freedom of design" when searching for small organic molecules which have a desired set of targeted QM properties, *i.e.*, molecules that are both highly polarizable and have high electrical stabilities as candidates for polymeric battery materials.

We then investigated how the QM7-X MPS depends on molecular structure and chemical composition (*i.e.*, the tunable "knobs" in molecular design) by considering the $(\langle E_{\text{MBD}}\rangle, \langle E_{\text{AT}}\rangle)$-space in more detail. Despite a moderate degree of correlation between these QM quantities, we still found that diverse $(\langle E_{\text{MBD}}\rangle, \langle E_{\text{AT}}\rangle)$ combinations are accessible, *i.e.*, we were able to easily identify molecules with very different structures (*i.e.*, compact vs. extended) and compositions with the same $\langle E_{\text{MBD}}\rangle$ but completely different $\langle E_{\text{AT}}\rangle$ and vice versa. Given such findings for two extensive QM energetic properties, we then looked for evidence of a similar "freedom of design" when targeting additional QM properties. In doing so, we found multiple cases where two distinct molecules shared four (extensive and intensive) global QM properties—a strong indication of the flexibility one has when designing molecules with a diverse and targeted array of QM properties. Local QM properties, which provide tremendous insight into the distinct chemical environments inside molecules, are largely uncorrelated with global QM properties and can instead be used as features to distinguish between molecules in high-dimensional MPS. As such, we argued that combinations of global and local QM properties (including extensive and intensive properties, as well as ground- and excited-state properties) can potentially be used to develop more robust molecular descriptors in ML applications. One could also imagine combining such rich QM-property-based molecular descriptors with current descriptors that only utilize structural information to further improve the transferability and scalability of next-generation ML models.

Since molecular design often involves the simultaneous optimization of multiple QM properties, we also used the extensive QM property data in QM7-X to explore more complex manifolds of MPS. From a tiered multi-property analysis of $(E_{\text{AT}}, E_{\text{gap}}, \alpha)$-space in which we partitioned the molecules in QM7-X according to an increasing number of properties, we again found a surprising degree of flexibility in MPS: by restricting the seemingly infinite search space to molecules with certain $E_{\text{AT}}$ (or certain $E_{\text{AT}}$ and $E_{\text{gap}}$), one can still find molecules with a markedly diverse range of complementary QM properties. Hence, we again found compelling evidence of the "freedom of design" that exists across wide swaths of a QM-based MPS, *i.e.*, the rational *in silico* design of molecules with completely different structures and compositions that share an array of different QM properties is quite feasible.

Based on these findings, we then employed Pareto front analysis, a powerful multi-property optimization approach, to identify the most promising small organic molecules in CCS (as enumerated by QM7-X) for polymeric battery materials (*i.e.*, molecules which simultaneously have the largest $\langle\alpha\rangle$ and $\langle E_{\text{gap}}\rangle$). In doing so, we found that the molecules in each Pareto front generally follow the inverse $\alpha \propto \frac{1}{E_{\text{gap}}}$ relationship; however, there were a number of exceptions and unexpected structures that appeared along each front, reflecting the freedom one has when designing molecules with multiple targeted QM properties. A deeper analysis of each front also revealed a series of small and rational changes to the structure and composition of each Pareto-optimal molecule that were very well-aligned with "chemical intuition". Since these findings emerged without any prior knowledge of the $(\alpha, E_{\text{gap}})$-space, we argued that Pareto front analysis is a powerful (and largely underutilized) tool for *in silico* molecular design[56–60]. A potentially interesting next step would use these Pareto-optimal structures in conjunction with current ML approaches (*e.g.*, active learning) to build reliable multi-objective frameworks for identifying the molecules in CCS (beyond that in QM7-X) missing in each front[63,64]. Such a framework would considerably improve the sampling, identification, and design of molecular systems for a number of applications, ranging from novel polymeric batteries and organic semiconductors to promising pharmaceuticals and small-molecule protein inhibitors. Hence, we hope that this work will emphasize the critical importance of obtaining high-quality QM property data and motivate the development of next-generation ML approaches that will allow us to gain a deeper and more fundamental understanding of the complex SPR/PPR in MPS as well as explore even more vast swaths of the seemingly infinite CCS—both of which are crucial for chemistry-based decision-making processes in the science, technology, and engineering fields.

# Methods

## Generation of the QM7-X dataset

In the construction of QM7-X[34], we performed a systematic and exhaustive sampling of the (meta-)stable equilibrium structures (*i.e.*, constitutional/structural isomers and stereoisomers, *e.g.*, enantiomers and diastereomers (including cis-/trans- and conformational isomers) of all molecules containing up to seven heavy (C, N, O, S, Cl) atoms in the GDB-13 database[24] using a density-functional tight binding (DFTB) approach[65,66] including many-body dispersion (MBD) interactions[35,67,68] for equilibrium structure generation. To further sample each molecular potential energy surface (PES), we generated 100 non-equilibrium conformers for each of these 41,357 equilibrium structures (*via* DFTB normal-mode displacements, see examples in Figure 1) producing a total of $\approx 4.2$ M molecular structures. For each of these equilibrium and non-equilibrium structures, QM7-X includes an extensive number of physicochemical properties (*i.e.*, 42 global (molecular), local (atom-in-molecule), ground-state, and response properties) obtained *via* QM calculations, most of which were computed using non-empirical hybrid density-functional theory (DFT) with a many-body treatment of vdW dispersion interactions (*i.e.*, PBE0+MBD) in conjunction with tightly-converged numeric atom-centered basis sets[69] as implemented in the `FHI-aims` code[70,71]. This level of theory has proven to be accurate and reliable for describing the intramolecular degrees of freedom in small organic molecules as well as the intermolecular interactions in organic molecular dimers, supramolecular complexes, and molecular crystals[35,72–75].

## Analysis details

For the analysis in the *Pairwise Correlations in Molecular Property Space* section, we considered the properties of all $\approx 4.2$ M (equilibrium and non-equilibrium) molecular structures in QM7-X. The degree of correlation between properties $X$ and $Y$ was measured by the Pearson correlation coefficient, *i.e.*,

$$\rho = \frac{\mathrm{cov}\,(X,Y)}{\sigma_X \sigma_Y} \,, \tag{3}$$

in which cov and $\sigma$ are the covariance and standard deviation, respectively. For the analysis in the *Multi-Property Analysis: Exploring More Complex Manifolds of Molecular Property Space* section, we considered the 51 lowest-energy non-equilibrium conformations per equilibrium structure ($\approx 2.1$ M structures, see energy range in Figure S4). The analyses in the *Structural and Compositional Dependence of Molecular Property Space* and *Multi-Property Optimization: Finding Optimal Pareto Fronts in Molecular Property Space* sections were performed using thermally-averaged values for each property at $T = 300$ K, *i.e.*, obtained by Boltzmann averaging over all 101 (equilibrium and non-equilibrium) molecular structures per equilibrium structure in QM7-X. We used thermal averages (represented by $\langle \cdots \rangle$ throughout this work) as this protocol is often employed in molecular design procedures.

## Multi-property optimization algorithm

Each Pareto front was found using a multi-objective evolutionary algorithm, *i.e.*, the non-dominated sorting genetic algorithm II (NSGA-II)[76,77], as implemented in the `pymoo` code[78]. NSGA-II performs a fast sorting of non-dominant samples to define the Pareto fronts, while the diversity in each front is controlled by a crowding-distance calculation[61,62]. In our proof-of-concept search for promising candidate molecules for polymeric battery materials, *i.e.*, molecules that are both highly polarizable and have high electrical stabilities, we employed the following two objective functions: $f_1(x) = x$ and $f_2(y) = y$, where $x = \langle E_{\mathrm{gap}} \rangle$ and $y = \langle \alpha \rangle$. Here, we note that the choice for these objective functions can be specifically tailored for a given application and modified accordingly, *e.g.*, $f(x) = x^2$ could be used for the molecular polarizability when looking for molecules with large vdW/dispersion interactions, given the quadratic $C_6 \propto \alpha^2$ relationship between these quantities (*cf.* Eq. (1)).

# Acknowledgement

LMS, JH, and AT acknowledge financial support from the European Research Council (ERC-CoG grant BeStMo). BGE and RAD acknowledge partial financial support from Cornell University through start-up funding. RAD also gratefully acknowledges financial support from an Alfred P. Sloan Research Fellowship. This research used resources of the Argonne Leadership Computing Facility, which is a DOE Office of Science User Facility supported under Contract DE-AC02-06CH11357. This research also used computational resources provided by the University of Luxembourg and the Center for Information Services and High Performance Computing (ZIH) at TU Dresden.

## Author contributions

The work was initially conceived by LMS and AT, and designed with contributions from JH, BGE, AVM, and RAD. AT and RAD supervised and revised all stages of the work. All authors discussed the results and contributed to the final manuscript.

## Data availability

The QM7-X dataset used in this work is available on https://doi.org/10.5281/zenodo.4288677.

## Competing interests

The authors declare no competing financial interests.

## Supplementary Information

- sup-info.pdf: Additional results supporting this analysis of the high-dimensional QM7-X molecular property space.

## References

1. Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O. & Walsh, A. Machine learning for molecular and materials science. *Nature* **559**, 547–555 (2018).
2. Huang, B. & von Lilienfeld, O. A. Ab initio machine learning in chemical compound space. *Chem. Rev.* **121**, 10001–10036 (2021).
3. von Lilienfeld, O. A., Müller, K.-R. & Tkatchenko, A. Exploring chemical compound space with quantum-based machine learning. *Nat. Rev. Chem.* **4**, 347–358 (2020).
4. Tkatchenko, A. Machine learning for chemical discovery. *Nat. Commun.* **11**, 4125 (2020).
5. Bartók, A. P. *et al.* Machine learning unifies the modeling of materials and molecules. *Sci. Adv.* **3**, 1–8 (2017).
6. Deringer, V. L. *et al.* Gaussian process regression for materials and molecules. *Chem. Rev.* **121**, 10073–10141 (2021).
7. Keith, J. A. *et al.* Combining machine learning and computational chemistry for predictive insights into chemical systems. *Chem. Rev.* **121**, 9816–9872 (2021).
8. Meuwly, M. Machine learning for chemical reactions. *Chem. Rev.* **121**, 10218–10239 (2021).
9. Muratov, E. N. *et al.* QSAR without borders. *Chem. Soc. Rev.* **49**, 3525–3564 (2020).
10. Cherkasov, A. *et al.* QSAR modeling: Where have you been? where are you going to? *J. Medicinal Chem.* **57**, 4977–5010 (2014).
11. Lambrinidis, G. & Tsantili-Kakoulidou, A. Challenges with multi-objective qsar in drug discovery. *Expert. Opin. on Drug Discov.* **13**, 851–859 (2018).
12. Clark, R. D. & Daga, P. R. *Building a Quantitative Structure-Property Relationship (QSPR) Model*, 139–159 (Springer New York, New York, NY, 2019).
13. Roy, K., Kar, S. & Das, R. *A Primer on QSAR/QSPR Modeling: Fundamental Concepts*. SpringerBriefs in Molecular Science (Springer International Publishing, 2015).
14. Le, T., Epa, V. C., Burden, F. R. & Winkler, D. A. Quantitative structure–property relationship modeling of diverse materials properties. *Chem. Rev.* **112**, 2889–2919 (2012).
15. Vamathevan, J. *et al.* Applications of machine learning in drug discovery and development. *Nat. Rev. Drug Discov.* **18**, 463–477 (2019).
16. Gawriljuk, V. O. *et al.* Development of machine learning models and the discovery of a new antiviral compound against yellow fever virus. *J. Chem. Inf. Model.* **61**, 3804–3813 (2021).
17. Stokes, J. M. *et al.* A deep learning approach to antibiotic discovery. *Cell* **180**, 688–702.e13 (2020).
18. Williams, T., McCullough, K. & Lauterbach, J. A. Enabling catalyst discovery through machine learning and high-throughput experimentation. *Chem. Mater.* **32**, 157–165 (2020).
19. Xue, D. *et al.* Accelerated search for materials with targeted properties by adaptive design. *Nat. Commun.* **7**, 11241 (2016).

**20.** Herbol, H. C., Hu, W., Frazier, P., Clancy, P. & Poloczek, M. Efficient search of compositional space for hybrid organic–inorganic perovskites via Bayesian optimization. *npj Comput. Mater.* **4**, 51 (2018).

**21.** Tallorin, L. *et al.* Discovering de novo peptide substrates for enzymes using machine learning. *Nat. Commun.* **9**, 5253 (2018).

**22.** Reymond, J.-L. & Awale, M. Exploring chemical space for drug discovery using the chemical universe database. *ACS Chem. Neurosci.* **3**, 649–657 (2012).

**23.** Ruddigkeit, L., van Deursen, R., Blum, L. C. & Reymond, J.-L. Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17. *J. Chem. Inf. Model.* **52**, 2864–2875 (2012).

**24.** Blum, L. C. & Reymond, J.-L. 970 million druglike small molecules for virtual screening in the chemical universe database GDB-13. *J. Am. Chem. Soc.* **131**, 8732–8733 (2009).

**25.** Fink, T., Bruggesser, H. & Reymond, J.-L. Virtual exploration of the small-molecule chemical universe below 160 Daltons. *Angew. Chem. Int. Ed.* **44**, 1504–1508 (2005).

**26.** Fink, T. & Reymond, J.-L. Virtual exploration of the chemical universe up to 11 atoms of C, N, O, F: Assembly of 26.4 million structures (110.9 million stereoisomers) and analysis for new ring systems, stereochemistry, physicochemical properties, compound classes, and drug discovery. *J. Chem. Inf. Model.* **47**, 342–353 (2007).

**27.** Montavon, G. *et al.* Machine learning of molecular electronic properties in chemical compound space. *New J. Phys.* **15**, 095003 (2013).

**28.** Rupp, M., Tkatchenko, A., Müller, K.-R. & von Lilienfeld, O. A. Fast and accurate modeling of molecular atomization energies with machine learning. *Phys. Rev. Lett.* **108**, 058301 (2012).

**29.** Ramakrishnan, R., Dral, P. O., Rupp, M. & von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Sci. Data* **1**, 140022 (2014).

**30.** Smith, J. S., Isayev, O. & Roitberg, A. E. ANI-1, A data set of 20 million calculated off-equilibrium conformations for organic molecules. *Sci. Data* **4**, 170193 (2017).

**31.** Smith, J. S., Isayev, O. & Roitberg, A. E. ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost. *Chem. Sci.* **8**, 3192–3203 (2017).

**32.** Yang, Y. *et al.* Quantum mechanical static dipole polarizabilities in the QM7b and AlphaML showcase databases. *Sci. Data* **6**, 152 (2019).

**33.** Smith, J. S. *et al.* The ANI-1ccx and ANI-1x data sets, coupled-cluster and density functional theory properties for molecules. *Sci. Data* **7**, 134 (2020).

**34.** Hoja, J. *et al.* QM7-X, a comprehensive dataset of quantum-mechanical properties spanning the chemical space of small organic molecules. *Sci. Data* **8**, 43 (2021).

**35.** Tkatchenko, A., DiStasio, R. A., Car, R. & Scheffler, M. Accurate and efficient method for many-body van der waals interactions. *Phys. Rev. Lett.* **108**, 236402 (2012).

**36.** DiStasio, R. A., von Lilienfeld, O. A. & Tkatchenko, A. Collective many-body van der waals interactions in molecular systems. *Proc. Natl. Acad. Sci.* **109**, 14791–14795 (2012).

**37.** Ambrosetti, A., Reilly, A. M., DiStasio, R. A. & Tkatchenko, A. Long-range correlation energy calculated from coupled atomic response functions. *J. Chem. Phys.* **140**, 18A508 (2014).

**38.** Al-Hamdani, Y. S. & Tkatchenko, A. Understanding non-covalent interactions in larger molecular complexes from first principles. *J. Chem. Phys.* **150**, 010901 (2019).

**39.** Brinck, T., Murray, J. S. & Politzer, P. Polarizability and volume. *J. Chem. Phys.* **98**, 4305–4306 (1993).

**40.** Blair, S. A. & Thakkar, A. J. Relating polarizability to volume, ionization energy, electronegativity, hardness, moments of momentum, and other molecular properties. *J. Chem. Phys.* **141**, 074306 (2014).

**41.** Casimir, H. B. G. & Polder, D. The influence of retardation on the London-van der Waals forces. *Phys. Rev.* **73**, 360–372 (1948).

**42.** Tang, K. T. & Karplus, M. Padé-approximant calculation of the nonretarded van der Waals coefficients for two and three helium atoms. *Phys. Rev.* **171**, 70–74 (1968).

**43.** Tkatchenko, A. & Scheffler, M. Accurate molecular van der Waals interactions from ground-state electron density and free-atom reference data. *Phys. Rev. Lett.* **102**, 073005 (2009).

**44.** DiStasio, R. A., Gobre, V. V. & Tkatchenko, A. Many-body van der Waals interactions in molecules and condensed matter. *J. Phys. Condens. Matter* **26**, 213202 (2014).

**45.** Tkatchenko, A., Ambrosetti, A. & DiStasio, R. A. Interatomic methods for the dispersion energy derived from the adiabatic connection fluctuation-dissipation theorem. *J. Chem. Phys.* **138**, 074106 (2013).

**46.** Brieger, M., Renn, A., Sodeik, A. & Hese, A. The dipole moment of 7LiH and 7LiD in the excited a $1\sigma+$ state: A test of the born-oppenheimer approximation. *Chem. Phys.* **75**, 1–9 (1983).

**47.** Buckingham, A. D. *Permanent and Induced Molecular Moments and Long-Range Intermolecular Forces*, vol. 12, chap. 2, 107–142 (John Wiley & Sons, Ltd, 1967).

**48.** Pouchan, C., Bégué, D. & Zhang, D. Y. Between geometry, stability, and polarizability: Density functional theory studies of silicon clusters $\mathrm{Si}_n$ (n=3-10). *J. Chem. Phys.* **121**, 4628–4634 (2004).

**49.** Meyers, F., Marder, S. R., Pierce, B. M. & Bredas, J. L. Electric field modulated nonlinear optical properties of donor-acceptor polyenes: Sum-over-states investigation of the relationship between molecular polarizabilities ($\alpha$, $\beta$, and $\gamma$) and bond length alternation. *J. Am. Chem. Soc.* **116**, 10703–10714 (1994).

**50.** Wilkins, D. M. *et al.* Accurate molecular polarizabilities with coupled cluster theory and machine learning. *Proc. Natl. Acad. Sci.* **116**, 3401–3406 (2019).

**51.** Fedorov, D. V., Sadhukhan, M., Stöhr, M. & Tkatchenko, A. Quantum-mechanical relation between atomic dipole polarizability and the van der Waals radius. *Phys. Rev. Lett.* **121**, 183401 (2018).

**52.** Lao, K. U., Yang, Y. & DiStasio, R. A. Electron confinement meet electron delocalization: non-additivity and finite-size effects in the polarizabilities and dispersion coefficients of the fullerenes. *Phys. Chem. Chem. Phys.* **23**, 5773–5779 (2021).

**53.** Yang, Y., Lao, K. U. & DiStasio, R. A. Influence of pore size on the van der Waals interaction in two-dimensional molecules and materials. *Phys. Rev. Lett.* **122**, 026001 (2019).

**54.** Yan, W. *et al.* All-polymer particulate slurry batteries. *Nat. Commun.* **10**, 2513 (2019).

**55.** Lopez, J., Mackanic, D. G., Cui, Y. & Bao, Z. Designing polymers for advanced battery chemistries. *Nat. Rev. Mater.* **4**, 312–330 (2019).

**56.** Farmahini, A. H., Krishnamurthy, S., Friedrich, D., Brandani, S. & Sarkisov, L. Performance-based screening of porous materials for carbon capture. *Chem. Rev.* **121**, 10666–10741 (2021).

**57.** Jablonka, K. M., Jothiappan, G. M., Wang, S., Smit, B. & Yoo, B. Bias free multiobjective active learning for materials design and discovery. *Nat. Commun.* **12**, 2312 (2021).

**58.** Gopakumar, A. M., Balachandran, P. V., Xue, D., Gubernatis, J. E. & Lookman, T. Multi-objective Optimization for Materials Discovery via Adaptive Design. *Sci. Reports* **8**, 3738 (2018).

**59.** Sun, Y. *et al.* Fingerprinting diverse nanoporous materials for optimal hydrogen storage conditions using meta-learning. *Sci. Adv.* **7**, eabg3983 (2021).

**60.** Erps, T. *et al.* Accelerated discovery of 3d printing materials using data-driven multiobjective optimization. *Sci. Adv.* **7**, eabf7435 (2021).

**61.** Mandal, J., Mukhopadhyay, S. & Dutta, P. *Multi-Objective Optimization: Evolutionary to Hybrid Framework* (Springer Singapore, 2018).

**62.** Rangaiah, G. & Bonilla-Petriciolet, A. *Multi-Objective Optimization in Chemical Engineering: Developments and Applications* (Wiley, 2013).

**63.** Janet, J. P., Ramesh, S., Duan, C. & Kulik, H. J. Accurate multiobjective design in a space of millions of transition metal complexes with neural-network-driven efficient global optimization. *ACS Cent. Sci.* **6**, 513–524 (2020).

**64.** del Rosario, Z., Rupp, M., Kim, Y., Antono, E. & Ling, J. Assessing the frontier: Active learning, model accuracy, and multi-objective candidate discovery and optimization. *The J. Chem. Phys.* **153**, 024112 (2020).

**65.** Seifert, G., Porezag, D. & Frauenheim, T. Calculations of molecules, clusters, and solids with a simplified LCAO-DFT-LDA scheme. *Int. J. Quantum Chem.* **58**, 185–192 (1996).

**66.** Gaus, M., Cui, Q. & Elstner, M. DFTB3: Extension of the self-consistent-charge density-functional tight-binding method (SCC-DFTB). *J. Chem. Theory Comput.* **7**, 931–948 (2011).

**67.** Stöhr, M., Michelitsch, G. S., Tully, J. C., Reuter, K. & Maurer, R. J. Communication: Charge-population based dispersion interactions for molecules and materials. *J. Chem. Phys.* **144**, 151101 (2016).

**68.** Mortazavi, M., Brandenburg, J. G., Maurer, R. J. & Tkatchenko, A. Structure and stability of molecular crystals with many-body dispersion-inclusive density functional tight binding. *J. Phys. Chem. Lett.* **9**, 399–405 (2018).

**69.** Havu, V., Blum, V., Havu, P. & Scheffler, M. Efficient $O(N)$ integration for all-electron electronic structure calculation using numeric basis functions. *J. Comput. Phys.* **228**, 8367–8379 (2009).

**70.** Blum, V. *et al.* Ab initio molecular simulations with numeric atom-centered orbitals. *Comput. Phys. Commun.* **180**, 2175 – 2196 (2009).

**71.** Ren, X. *et al.* Resolution-of-identity approach to hartree–fock, hybrid density functionals, RPA, MP2 and GW with numeric atom-centered orbital basis functions. *New J. Phys.* **14**, 053020 (2012).

**72.** Perdew, J. P., Ernzerhof, M. & Burke, K. Rationale for mixing exact exchange with density functional approximations. *J. Chem. Phys.* **105**, 9982–9985 (1996).

**73.** Adamo, C. & Barone, V. Toward reliable density functional methods without adjustable parameters: The PBE0 model. *J. Chem. Phys.* **110**, 6158–6170 (1999).

**74.** Lynch, B. J. & Truhlar, D. G. Robust and affordable multicoefficient methods for thermochemistry and thermochemical kinetics: the MCCM/3 suite and SAC/3. *J. Phys. Chem. A* **107**, 3898–3906 (2003).

**75.** Hoja, J. *et al.* Reliable and practical computational description of molecular crystal polymorphs. *Sci. Adv.* **5**, eaau3338 (2019).

**76.** Deb, K., Pratap, A., Agarwal, S. & Meyarivan, T. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE T. Evol. Comput.* **6**, 182–197 (2002).

**77.** Bhaskar, V., Gupta, S. K. & Ray, A. K. Applications of multiobjective optimization in chemical engineering. *Rev. Chem. Eng.* **16**, 1 – 54 (01 Mar. 2000).

**78.** Blank, J. & Deb, K. Pymoo: Multi-objective optimization in Python. *IEEE Access* **8**, 89497–89509 (2020).