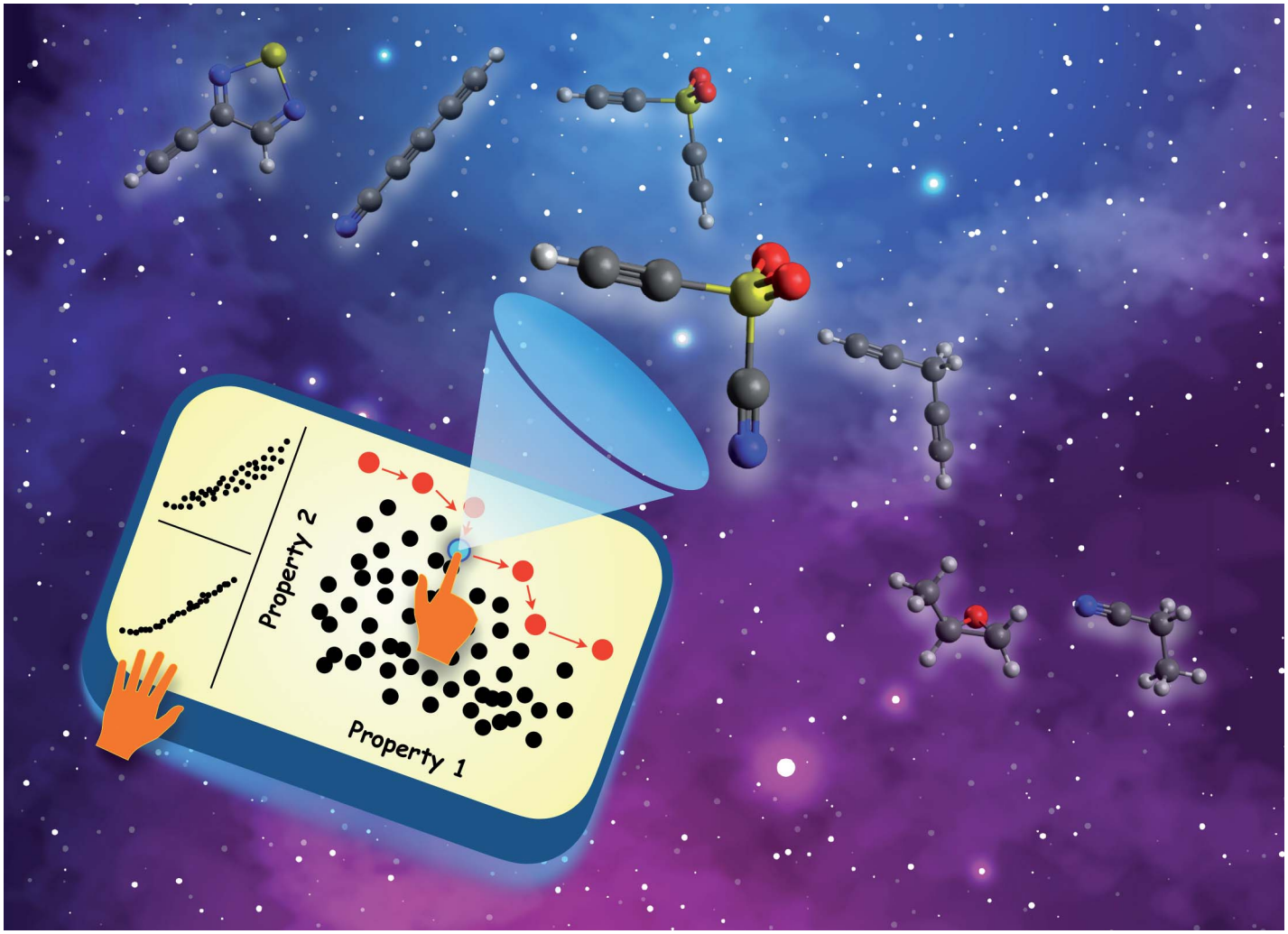

Showcasing research from Professor Alexandre Tkatchenko's group, Department of Physics and Materials Science, University of Luxembourg, Luxembourg City, Luxembourg.

"Freedom of design" in chemical compound space: towards rational *in silico* design of molecules with targeted quantum-mechanical properties

This work demonstrates that "freedom of design" is a fundamental and emergent property of chemical compound space – the unfathomably vast space populated by all possible atomic compositions and their geometries. Such intrinsic flexibility enables rational design of distinct molecules sharing an array of targeted quantum-mechanical properties. The combination of the insights gained from this work with advanced machine learning approaches could aid in the development of effective strategies for high-throughput screening of novel molecules tailored to a specific application. Galaxy background image by pikisuperstar *via* Freepik.

As featured in:

See Leonardo Medrano Sandonas, Robert A. DiStasio, Jr, Alexandre Tkatchenko *et al.*, *Chem. Sci.*, 2023, **14**, 10702.

ROYAL SOCIETY OF CHEMISTRY

rsc.li/chemical-science

Registered charity number: 207890

Chemical Science

ROYAL SOCIETY OF CHEMISTRY

EDGE ARTICLE

View Article Online
View Journal | View Issue

Cite this: *Chem. Sci.*, 2023, **14**, 10702

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

"Freedom of design" in chemical compound space: towards rational *in silico* design of molecules with targeted quantum-mechanical properties†

Leonardo Medrano Sandonas, *[a] Johannes Hoja, [ab] Brian G. Ernst, [c] Álvaro Vázquez-Mayagoitia, [d] Robert A. DiStasio, Jr *[c] and Alexandre Tkatchenko *[a]

Received 13th July 2023
Accepted 17th August 2023

DOI: 10.1039/d3sc03598k

rsc.li/chemical-science

The rational design of molecules with targeted quantum-mechanical (QM) properties requires an advanced understanding of the structure–property/property–property relationships (SPR/PPR) that exist across chemical compound space (CCS). In this work, we analyze these fundamental relationships in the sector of CCS spanned by small (primarily organic) molecules using the recently developed QM7-X dataset, a systematic, extensive, and tightly converged collection of 42 QM properties corresponding to ≈4.2M equilibrium and non-equilibrium molecular structures containing up to seven heavy/non-hydrogen atoms (including C, N, O, S, and Cl). By characterizing and enumerating progressively more complex manifolds of molecular property space—the corresponding high-dimensional space defined by the properties of each molecule in this sector of CCS—our analysis reveals that one has a substantial degree of flexibility or "freedom of design" when searching for a single molecule with a desired pair of properties or a set of distinct molecules sharing an array of properties. To explore how this intrinsic flexibility manifests in the molecular design process, we used multi-objective optimization to search for molecules with simultaneously large polarizabilities and HOMO–LUMO gaps; analysis of the resulting Pareto fronts identified non-trivial paths through CCS consisting of sequential structural and/or compositional changes that yield molecules with optimal combinations of these properties.

1 Introduction

In recent years, exploration of the remarkably vast chemical compound space (CCS) of molecules and materials with data-driven approaches has inspired countless academic and industrial initiatives to seek out the fundamental relationships that exist (among and) between the structural signatures of molecules (*e.g.*, chemical composition, atom connectivity, molecular structure) and their physicochemical properties (*e.g.*, energies, HOMO–LUMO gaps, polarizabilities).[1–8] In doing so, the growing availability of accurate and reliable molecular property data, coupled with the use of machine learning (ML) algorithms to explore this data, have led to an increased qualitative and quantitative understanding of these structure–property/property–property relationships (SPR/PPR).[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]—scientific and technological endeavors that have traditionally been driven by chemical intuition and/or serendipitous discoveries. While there has been significant progress in this area, a comprehensive understanding of these complex SPR/PPR—even in the (relatively) more manageable sector of CCS spanned by small molecules—is still lacking despite the critical importance and high relevance of such molecules throughout the chemical sciences. Unravelling these complex relationships would not only provide us with the tools needed to explore and characterize molecular property space (*i.e.*, the even higher dimensional space defined by the properties of each molecule in CCS), but would also greatly advance our ability to rationally design molecules with targeted arrays of physicochemical properties.

To address this challenge, the GDB databases[22–26] have enumerated the molecular graphs for a sizeable number of chemical compounds (≈166G)—an important first step towards the systematic exploration of swaths of CCS far too vast to be cataloged and studied experimentally. Since these graphs

[a]*Department of Physics and Materials Science, University of Luxembourg, L-1511 Luxembourg City, Luxembourg. E-mail: leonardo.medrano@uni.lu; alexandre.tkatchenko@uni.lu*

[b]*Institute of Chemistry, University of Graz, 8010 Graz, Austria*

[c]*Department of Chemistry and Chemical Biology, Cornell University, Ithaca, NY 14853, USA. E-mail: distasio@cornell.edu*

[d]*Computational Science Division, Argonne National Laboratory, Lemont, IL 60439, USA*

† Electronic supplementary information (ESI) available: Additional analyses of the structure–property and property–property relationships in the sector of chemical compound space spanned by small (primarily organic) molecules. See DOI: https://doi.org/10.1039/d3sc03598k

only contain chemical composition (*i.e.*, molecular formula) and atom connectivity information, the three-dimensional (3D) molecular structure(s) consistent with each graph (as well as their corresponding physicochemical properties) still need to be determined before detailed SPR/PPR studies can be undertaken. Furthermore, while certain chemical rules/guidelines were used to generate graphs corresponding to stable (and potentially synthetically feasible) molecules, there will be graphs containing atom-connectivity motifs that are more prone to stability issues when translated into 3D molecular structures (*e.g.*, small rings with high ring strain). To address these issues, several researchers have built upon these seminal databases by computing quantum-mechanical (QM) structure and property information for the subset of molecular graphs containing $\lesssim$10 heavy/non-hydrogen atoms, *i.e.*, the graphs enumerating the sector of CCS spanned by small (primarily organic) molecules mentioned above.[27–34] For instance, the QM7 dataset[27,28] provides the equilibrium structures and 15 physicochemical properties for 7211 molecules corresponding to the molecular graphs from GDB-13 (ref. 24) that contain up to seven heavy/non-hydrogen atoms (including C, N, O, S, and Cl); such structural and property information was computed using a hierarchy of different QM methods (*i.e.*, ZINDO, SCS, PBE0, GW), with more recent variants employing (LR-)CCSD for molecular property evaluation.[32] The subsequent QM9 dataset[29] generated the structures and 16 (geometric, energetic, electronic, and thermodynamic) properties for 133 885 molecules at the B3LYP/6-31G(2df,p) level, each of which corresponds to a molecular graph from GDB-17 (ref. 23) containing up to nine heavy atoms (including C, N, O, and F). Several years later, another extensive exploration of the small-molecule sector of CCS was accomplished by the ANI-1 dataset,[30,31] which consists of more than 20M equilibrium and non-equilibrium molecular structures containing up to eight heavy atoms (albeit limited to C, N, and O only) based on molecular graphs from GDB-11.[25,26] This was followed by the release of the ANI-1x and ANI-1ccx datasets,[33] which contain 20 properties for $\approx$5M structures computed using the ωB97-X functional and energies of $\approx$5k structures computed at the CCSD(T)/CBS level, respectively.

Despite all of these foundational efforts to generate a fully QM description of the sector of CCS spanned by small molecules (and the corresponding sector of molecular property space), many challenges still exist when translating a series of molecular graphs (which only contain chemical composition and atom connectivity information) to a systematic sampling of such high-dimensional spaces that includes an accurate and reliable account of both structural information (*i.e.*, equilibrium and non-equilibrium structures consistent with each molecular graph) and property information (*i.e.*, an extensive and well-converged inventory of QM properties for each molecular structure). To address these challenges, the recently published QM7-X dataset[34] provides a systematic, extensive, and tightly converged collection of 42 QM properties for $\approx$4.2M equilibrium and non-equilibrium structures corresponding to the $\approx$7k molecular graphs containing up to seven heavy atoms (C, N, O, S, Cl) in the GDB-13 database,[24] providing what is arguably the most comprehensive QM description of the sectors of CCS *and* molecular property space spanned by small (primarily organic) molecules to date. Rather than just including a single molecular structure per graph, QM7-X contains an extensive sampling of $\approx$42k (meta-)stable equilibrium structures/isomers corresponding to this set of molecular graphs, as well as 100 additional non-equilibrium conformations per equilibrium structure. For each of the resulting $\approx$4.2M equilibrium and non-equilibrium molecular structures, QM7-X also includes an extensive number of physicochemical properties (*i.e.*, 42 structural, global (molecular), local (atom-in-molecule), ground-state, and response properties) obtained using well-converged and highly accurate QM methodologies, thereby providing the prerequisite QM description of both structural and property information needed for impactful SPR/PPR research efforts.

In this work, we build upon this foundational effort by performing a detailed analysis of the molecular property space corresponding to small (primarily organic) molecules (*i.e.*, as enumerated by the QM7-X dataset) in order to gain a deeper understanding of the SPR/PPR that exist throughout this sector of CCS. We start with a quantitative analysis of the pairwise correlations between select QM7-X properties, which revealed that most (>90%) properties exhibit little (to no) correlation, *i.e.*, there are very few strict limitations preventing a single molecule from simultaneously exhibiting a desired pair of QM properties. By progressively investigating more complex manifolds of the QM7-X molecular property space and their underlying dependence on molecular structure and chemical composition (*i.e.*, the tunable "knobs" in molecular design), we also found a remarkably large number of structurally and/or compositionally distinct molecules that share multiple QM properties. Taken together, these findings provide compelling evidence for "freedom of design" in CCS—an intrinsic degree of flexibility that enables the rational design of distinct molecules sharing an array of targeted physicochemical properties. To explore how this intrinsic flexibility manifests in the molecular design process, we used Pareto multi-property optimization to search for molecules in QM7-X with simultaneously large polarizabilities (α) and HOMO–LUMO gaps (E_{gap}). Without any prior knowledge of the corresponding (α, E_{gap})-manifold of molecular property space, each Pareto front reflected the large degree of intrinsic flexibility within CCS by identifying a series of changes to the molecular structure and/or chemical composition that resulted in optimal combinations of these properties.

By demonstrating that "freedom of design" is a fundamental and emergent property of CCS, we hope this work will challenge the greater chemical sciences community to consider how such intrinsic flexibility can be used to expand the candidate pool of chemical compounds—beyond the paradigm of functional group modification based on a largely fixed molecular scaffold(s)—during the rational design of molecules with targeted physicochemical properties. We expect that the insight provided by this work will emphasize the critical importance of obtaining high-quality QM structural and property data when exploring the fundamental SPR/PPR existing throughout CCS and contribute to the development of advanced ML-based tools that will improve the *in silico* 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 Methods

2.1 Description of the QM7-X dataset

The QM7-X dataset[34] was constructed *via* a systematic and extensive 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)) corresponding to the ≈7k molecular graphs containing up to seven heavy/non-hydrogen atoms (including C, N, O, S, and Cl) in the GDB-13 database.[24] As mentioned above, these molecular graphs only contain chemical composition and atom connectivity information, and therefore do not specify the 3D molecular structures (and their corresponding physicochemical properties) that are consistent with each graph. Furthermore, there will be graphs that contain certain moieties (*e.g.*, small rings) that are more prone to stability issues (*e.g.*, high ring strain) when translated into fully specified 3D molecular structures. To address these issues, we optimized each of the molecular structures in QM7-X using accurate and reliable QM methodologies, *i.e.*, third-order self-consistent charge density-functional tight binding (DFTB3)[35,36] that also accounts for many-body dispersion/van der Waals (vdW) interactions *via* the MBD approach,[37,38] and performed subsequent harmonic frequency analyses on the resulting set of 41 537 optimized structures to confirm that each was a stationary point (*i.e.*, a local minimum) on the corresponding molecular potential energy surface (PES). We also confirmed that each of these optimized molecular structures has a positive atomization energy ($E_{AT} > 0$), and is therefore energetically downhill with respect to the infinite-separation limit of its constituent atoms (*cf.* Fig. 4(a) in ref. 34). While this protocol cannot guarantee that each molecular structure in QM7-X is stable with respect to the (Gibbs) free energy, this set of structures was taken to be a representative sample of the (meta-)stable equilibrium molecules in the sector of CCS spanned by small primarily organic molecules.

To further sample each molecular PES, we also generated 100 non-equilibrium conformations for each of these ≈42k equilibrium structures (*via* the normal-mode displacements obtained during the harmonic frequency analysis, see schematic illustration in Fig. 1), yielding a total of ≈4.2M molecular structures. We note in passing that each of these non-equilibrium conformations also has a positive atomization energy (*cf.* Fig. 4(a) in ref. 34), despite the energetic destabilization resulting from the generative structural perturbations. For each of these ≈4.2M equilibrium and non-equilibrium structures, QM7-X also includes an extensive number of physicochemical properties (*i.e.*, 42 structural, global (molecular), local (atom-in-molecule), ground-state, and response properties) obtained from QM calculations, most of which were performed with non-empirical hybrid density-functional theory (DFT) and a many-body treatment of vdW/dispersion interactions (*i.e.*, PBE0+MBD)[37,39–41] in conjunction with the tightly converged numeric atom-centered basis sets[42] implemented in the `FHI-aims` code.[43,44]

2.2 Analysis of the QM7-X molecular property space

Our detailed analysis of the QM7-X molecular property space includes the following four thrusts: (i) projecting the corresponding 42D molecular property space onto a series of 2D correlation plots for identifying pairwise PPR (Sec. 3.1); (ii) characterizing the structural and compositional dependence of select QM7-X properties (Sec. 3.2); (iii) quantifying the number of structurally and/or compositionally distinct molecules that share an array of QM7-X properties (Sec. 3.3); (iv) finding and analyzing Pareto fronts containing molecules in QM7-X with simultaneously large polarizabilities (α) and HOMO–LUMO gaps (E_{gap}) (Sec. 3.4).

For the analysis in Sec. 3.1, we considered the properties of all ≈4.2M (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,

$$\rho_{x,y} = \frac{\mathrm{cov}(x,y)}{\sigma_x \sigma_y}, \quad (1)$$

in which cov(x,y) and $\sigma_{x/y}$ are the covariance and standard deviation, respectively. The analyses in Secs. 3.2, 3.3, and 3.4 were performed using thermally-averaged values for each property at $T = 300$ K, which were obtained by Boltzmann averaging over all 101 (equilibrium and non-equilibrium) conformations per equilibrium structure in QM7-X. Thermal averages (denoted by $\langle\cdots\rangle$ throughout) were specifically used in these sections as this convention is commonly employed in molecular design protocols. To facilitate the analysis in Sec. 3.2, the spatial extent of each molecule was accounted for using D_{max}, the *maximum* distance between pairs of heavy/non-hydrogen atoms in a given molecular structure. To gain additional insight into the SPR/PPR involving extensive properties in Sec. 3.2, we also considered normalized variants of these properties (denoted by a prime superscript throughout). In such cases, the extensive properties were normalized with respect to the following two quantities: (i) the thermally-averaged molecular volume $V = \frac{4}{3}\pi\langle R_g\rangle^3$ (in which $\langle R_g\rangle$ is the thermally-averaged radius of gyration at $T = 300$ K); and (ii) the total number of atoms N_{atoms} in a given molecule. In Sec. 3.3, we used the following property-specific threshold values (each of which represents 0.5% of the total range observed for each thermally-averaged property at $T = 300$ K in the QM7-X dataset) when identifying a set of molecules sharing a given property value: $\delta_{\langle E_{AT}\rangle} = 0.42$ eV for the thermally-averaged atomization energy $\langle E_{AT}\rangle$, $\delta_{\langle\alpha\rangle} = 0.58\ a_0^3$ for the thermally-averaged isotropic molecular polarizability $\langle\alpha\rangle$, $\delta_{\langle E_{gap}\rangle} = 0.04$ eV for the thermally-averaged HOMO–LUMO gap $\langle E_{gap}\rangle$, and $\delta_{\langle\mu\rangle} = 0.008\ e\cdot$Å for the thermally-averaged scalar molecular dipole moment $\langle\mu\rangle$. For example, a three-molecule set that shares the same $\langle E_{AT}\rangle$ value will be composed of molecules A, B, and C provided that the following three conditions are satisfied: $|\langle E_{AT}\rangle_A - \langle E_{AT}\rangle_B| < \delta_{\langle E_{AT}\rangle}$, $|\langle E_{AT}\rangle_A - \langle E_{AT}\rangle_C| < \delta_{\langle E_{AT}\rangle}$, and $|\langle E_{AT}\rangle_B - \langle E_{AT}\rangle_C| < \delta_{\langle E_{AT}\rangle}$.

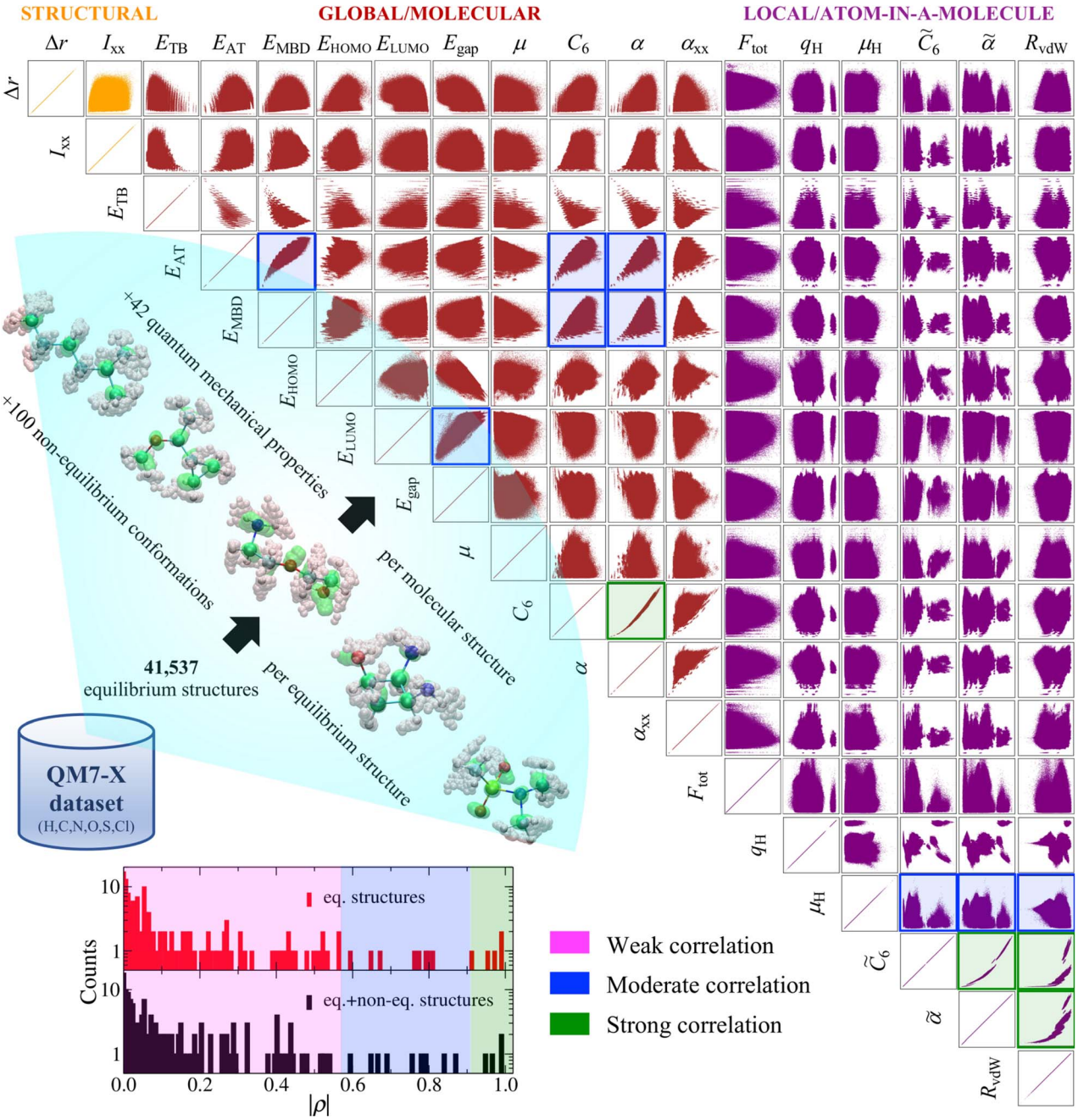

Fig. 1 Pairwise property–property relationships (PPR) in molecular property space: 2D correlation plots. The recently developed QM7-X dataset[34]—a systematic, extensive, and tightly converged collection of 42 quantum mechanical (QM) properties corresponding to ≈4.2M equilibrium and non-equilibrium molecular structures containing up to seven heavy/non-hydrogen atoms (including C, N, O, S, and Cl)—is used in this work to study the PPR in the sector of chemical compound space (CCS) spanned by small (primarily organic) molecules. Select 2D projections of the 42D QM7-X molecular property space are depicted for a subset of 18 structural (orange), global/molecular (red), and local/atom-in-a-molecule (violet) properties (see Table 1 for a detailed description of each property) for the ≈4.2M structures in QM7-X. As depicted in the *lower-left inset*, measuring the degree of correlation between each pair of QM properties (*via* the Pearson correlation coefficient ρ in eqn (1)) results in three distinct clusters: weakly correlated ($|\rho| \leq 0.57$), moderately correlated ($0.57 < |\rho| \leq 0.91$, highlighted with blue frames), and strongly correlated ($|\rho| > 0.91$, highlighted with dark green frames). A vast majority (*i.e.*, 140/153 or 91.5%) of these correlation plots resemble structureless "blobs" (*i.e.*, weakly correlated PPR with $|\rho| \leq 0.57$), indicating that small (primarily organic) molecules have the flexibility to exhibit nearly any pair of properties considered above.

2.3 Multi-property optimization algorithm

Each Pareto front in Sec. 3.4 was found using a multi-objective evolutionary algorithm, *i.e.*, the non-dominated sorting genetic algorithm II (NSGA-II),[45,46] as implemented in the `pymoo` code.[47] 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.[48,49] In our search for molecules in QM7-X with simultaneously large $\langle\alpha\rangle$

and $\langle E_{gap} \rangle$ values, we employed the following two objective functions: $f_1(x) = x$ and $f_2(y) = y$, in which $x = \langle E_{gap} \rangle$ and $y = \langle \alpha \rangle$.[50,51] Here, we note in passing that these objective functions could also be tailored for other potential applications, *i.e.*, $f(x) = x^2$ could be employed for α when searching for molecules with large vdW/dispersion interactions, given the quadratic relationship between α and the isotropic molecular vdW/dispersion coefficient C_6 (*cf.* $C_6 \propto \alpha^2$ in eqn (2)).

3 Results & discussion

3.1 Pairwise correlations in molecular property space

As a first step towards gaining a deeper understanding of the structure–property/property–property relationships (SPR/PPR) in the sector of chemical compound space (CCS) spanned by small (primarily organic) molecules, we analyzed the pairwise correlations between select properties in the QM7-X dataset. To do so, we plotted an array of 2D projections of the 42D QM7-X molecular property space corresponding to 18 QM properties evaluated on the ≈4.2M (equilibrium and non-equilibrium) structures in Fig. 1. This set of QM7-X properties is quite diverse and includes both extensive and intensive properties, as well as representative examples of structural, global/molecular, local/atom-in-a-molecule, ground-state, and response properties (see Table 1). Even from a cursory glance at Fig. 1, one can see that nearly all of the 153 unique pair projections (*i.e.*, 2D correlation plots) resemble structureless "blobs", indicating that most of these QM properties are effectively uncorrelated. To quantify the degree of correlation in each plot, we computed the corresponding Pearson correlation coefficient in eqn (1), and found three distinct and fairly well-defined groups of ρ values (see Fig. 1 inset). For the purposes of this discussion, we will use these values to (roughly) classify a given PPR as strongly correlated ($|\rho| > 0.91$), moderately correlated ($0.57 < |\rho| \leq 0.91$), or weakly correlated ($|\rho| \leq 0.57$). Under this working classification system, a mere 4/153 (2.6%) projections displayed a strong degree of correlation with $|\rho| > 0.91$: (C_6, α), ($\tilde{C}_6$, $\tilde{\alpha}$), ($\tilde{C}_6$, R_{vdW}), and ($\tilde{\alpha}$, R_{vdW}); these correlation plots are highlighted with dark green frames in Fig. 1. In the same breath, only 9/153 (5.9%) projections were classified as moderately correlated, with intermediate $|\rho|$ values and considerably more dispersion in their scatter plots: (E_{AT}, E_{MBD}), (E_{AT}, C_6), (E_{AT}, α), (E_{MBD}, C_6), (E_{MBD}, α), (E_{LUMO}, E_{gap}), (μ_H,$\tilde{C}_6$), (μ_H,$\tilde{\alpha}$), and (μ_H, R_{vdW}); these plots are highlighted with blue frames in Fig. 1. Hence, the remaining 140/153 (91.5%)—the lion's share of the correlation plots in Fig. 1—correspond to weakly correlated PPR with $|\rho| \leq 0.57$, which are particularly low values in the physical sciences. In other words, most of these properties are effectively uncorrelated, and there seem to be very few limitations preventing a single molecule from simultaneously exhibiting any pair of properties in Fig. 1. This finding is remarkable and implies that the sector of CCS spanned by small (primarily organic) molecules has a large degree of intrinsic flexibility—an important point that we return to throughout this manuscript.

Among the 2D projections in Fig. 1 that do exhibit a strong (or moderate) degree of correlation, we observed several trends that can be explained using chemical/physical intuition. For instance, consider the strong degree of correlation ($|\rho| = 0.99$) between the isotropic molecular vdW/dispersion coefficient (C_6) and isotropic molecular polarizability (α). In this case, the observed quadratic form can be rationalized by the Casimir–Polder formula[52] for the C_6 coefficient describing the vdW/dispersion interactions between molecules A and B:

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

in which $\hat{\alpha}_A(i\omega)$ is the frequency-dependent isotropic polarizability of A in the imaginary frequency domain. By substituting the leading-order Padé[53,54] (or quantum harmonic oscillator)[55,56] approximation for $\hat{\alpha}_A(i\omega)$ into this expression (*i.e.*, $\hat{\alpha}_A(i\omega) = \alpha_A/[1 - (\omega/\eta_A)^2]$, with η_A being the characteristic excitation frequency of A), one arrives at the well-known London formula on the right hand side of eqn (2) in which $C_6 \propto \alpha^2$.[57] Since the leading-order term in the many-body expansion of the vdW/dispersion energy is given by C_6/R^6,[57] it is also not surprising to observe a moderate degree of correlation ($|\rho| = 0.65$) between the many-body vdW/dispersion energy[37,41,58,59] (E_{MBD}) and C_6, as well as a similar degree of correlation ($|\rho| = 0.59$) between E_{MBD} and α (since $C_6 \propto \alpha^2$). Another related example is the moderate correlation observed between α_{xx} (a single component of the molecular polarizability tensor) and C_6 ($|\rho| = 0.52$); this is significantly less than the strong correlation ($|\rho| = 0.99$) between C_6 and α (the isotropic average over the diagonal tensor components), and directly reflects the anisotropy in the shapes and spatial extents of the molecules in QM7-X. Here, we stress that such a reduced degree of correlation between fundamental QM properties is by no means uninteresting, and reflects a degree of flexibility that can be exploited when searching for molecules with specific properties, *e.g.*, polarization directions/orientations that favor (or disfavor) the formation of specific molecular crystal polymorphs. If one considers the 2D projection between the atomization energy (E_{AT}) and E_{MBD}, one can (at least partially) rationalize the observed moderate degree of correlation ($|\rho| = 0.87$) by recognizing that E_{AT} and E_{MBD} both depend on molecular size; in this case, our physical intuition regarding extensive properties tells us that E_{AT} will *tend* to increase with E_{MBD}, and *vice versa*. In the same breath, the moderate degree of correlation ($|\rho| = 0.67$) between E_{AT} and α can also be rationalized with the physical/chemical intuition that α is (in general) additive and tends to increase with molecular volume.[60,61] As discussed below, one can normalize such extensive properties with respect to a number of different size-dependent quantities (*e.g.*, molecular volume, total number of atoms, number of valence electrons); however, the resulting degree of correlation will depend on the normalization quantity and can therefore be quite variable and counterintuitive (see Sec. 3.2 and Fig. S4†).

Using the extensive structure–property data in QM7-X, Fig. 1 also provides new insight into the correlations (or lack thereof) that exist between fundamental QM properties. At the top of this list is the seemingly expected inverse proportionality between α and HOMO–LUMO gap (E_{gap}),[61–73] which has roots in the following sum-over-states expression for α from perturbation theory:[74,75]

Table 1 Pairwise property–property relationships (PPR) in molecular property space: quantum mechanical (QM) properties. List of QM properties (and corresponding symbols) taken from the QM7-X dataset[34] and considered in the pairwise PPR analysis in Fig. 1. In the units and dimension provided for each of these QM properties, E_h and a_0 represent the atomic units of energy (Hartree) and length (Bohr radius), respectively, and N_{atoms} is the total number of atoms in a given molecular structure. Property types and classes were categorized as follows: structural (S), global/molecular (M), local/atom-in-a-molecule (A), ground-state (G), response (R), extensive (E), and intensive (I)

Symbol	Property description	Units	Dimension	Type	Class
Δr	RMSD with respect to equilibrium structure	Å	1	S,G	I
I_{xx}	Moment of inertia tensor (xx component)	amu·Å^2	1	S,G	I
D_{max}	Maximum distance between heavy/non-hydrogen atoms	Å	1	S,G	I
E_{TB}	Total DFTB energy	eV	1	M,G	E
E_{AT}	Atomization energy	eV	1	M,G	E
E_{MBD}	MBD energy	eV	1	M,G	E
E_{HOMO}	HOMO energy	eV	1	M,G	I
E_{LUMO}	LUMO energy	eV	1	M,G	I
E_{gap}	HOMO–LUMO gap	eV	1	M,G	I
μ	Scalar molecular dipole moment	e·Å	1	M,G	I
C_6	Isotropic molecular vdW/dispersion coefficient	$E_h \cdot a_0^6$	1	M,R	E
α	Isotropic molecular polarizability	a_0^3	1	M,R	E
α_{xx}	Molecular polarizability tensor (xx component)	a_0^3	1	M,R	E
F_{tot}	Norm of total atomic force vector	eV·Å^{-1}	1	A,G	I
q_H	Hirshfeld atomic charges	e	N_{atoms}	A,G	I
μ_H	Scalar Hirshfeld atomic dipole moments	$e \cdot a_0$	N_{atoms}	A,G	I
$\tilde{C}_6$	Isotropic atomic vdW/dispersion coefficients	$E_h \cdot a_0^6$	N_{atoms}	A,R	I
$\tilde{\alpha}$	Isotropic atomic polarizabilities	a_0^3	N_{atoms}	A,R	I
R_{vdW}	Isotropic atomic van der Waals (vdW) radii	a_0	N_{atoms}	A,R	I

$$\alpha = 2\sum_{k\neq 0}\frac{|\langle 0|\mu|k\rangle|^2}{E_k - E_0} \approx \frac{|\langle \mathrm{HOMO}|\mu|\mathrm{LUMO}\rangle|^2}{E_{gap}}, \quad (3)$$

in which $\langle 0|$ and $|k\rangle$ are the ground-state and excited-state electronic wavefunctions, E_0 and E_k are the associated eigenenergies, and $\langle 0|\mu|k\rangle$ is the corresponding transition dipole moment matrix element. When evaluating eqn (3) for molecules using a mean-field one-electron theory (*e.g.*, Hartree–Fock or Kohn–Sham density functional theory), it is often assumed that the lowest-energy HOMO → LUMO transition makes the most significant contribution to the summation above (as long as it is symmetry allowed). Subsequent approximation of this sum by this single term leads to the expression on the right hand side of eqn (3), in which $\alpha \propto \frac{1}{E_{gap}}$. Notably, this inverse proportionality also results from analytical evaluation of this sum-over-states expression for the quantum harmonic oscillator,[65,66] which is effectively transformed into a two-state system by the selection rules. Furthermore, we note that this relationship has also become part of physical/chemical intuition, primarily through HSAB (hard–soft–acid–base) theory[67,68] and the closely related concept of chemical hardness,[61,69,72] both of which are often evoked to rationalize the stability/reactivity of chemical species; this relationship also appears *via* connections between α (and/or E_{gap}) and a number of different theoretical quantities (*e.g.*, ionization energy, electron affinity, electronegativity).[61,70] While such an inverse proportionality has certainly been observed in homologous sets of molecules (*e.g.*, polyenes[71] and *s-trans* alkenes[73] of increasing length), we find that this relationship does not hold for the diverse set of molecules in QM7-X. As depicted in Fig. 1, the correlation plot between α and E_{gap} has only the faintest indication of an inverse proportionality and is more appropriately described as a structureless "blob"; with a correlation coefficient of $|\rho| = 0.06$ (which is abysmally low in the physical sciences), we would argue that these properties are effectively uncorrelated.

Unlike the majority of global/molecular properties (which have correlation plots resembling single connected "blobs"), the 2D projections involving local/atom-in-a-molecule properties, *e.g.*, Hirshfeld atomic charges (q_H), scalar Hirshfeld atomic dipole moments (μ_H), isotropic atomic vdW/dispersion coefficients ($\tilde{C}_6$), and isotropic atomic polarizabilities ($\tilde{\alpha}$), often exhibit distinct clusters. As depicted in Fig. 1, such clusters are most visible when analyzing the correlation plots 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_H show the largest number of atomic environments, and represent the different local charge distributions that exist throughout this diverse dataset. Local response properties such as $\tilde{C}_6$, $\tilde{\alpha}$, and R_{vdW} (*i.e.*, isotropic atomic vdW radii, see Table 1) also depend on the chemical environment surrounding each atom and tend to be strongly

correlated. For instance, one can observe multiple quadratic-type functions in the ($\tilde{C}_6$, $\tilde{\alpha}$)-plot, which can be rationalized by applying the Casimir–Polder relationship in eqn (2) to each local chemical environment. In the same breath, we also find a strong degree of correlation between $\tilde{\alpha}$ and R_{vdW} ($|\rho| = 0.97$)—this is a fundamental relationship that continues to be a topic of discussion in the literature.[76,77]

The distinct clustering of $|\rho|$ values discussed above is quite robust and also holds when only considering the ≈41k equilibrium structures in QM7-X, *i.e.*, the inclusion of the remaining ≈4.2M non-equilibrium conformations in QM7-X does not meaningfully alter the classification scheme used to discuss the PPR in this work (see the lower-left inset in Fig. 1 and S1† for representative examples). Statistically speaking, the mean absolute deviation (MAD) between $|\rho|$ values in the two distributions depicted in the Fig. 1 inset is quite small (MAD = 0.04), and is primarily due to inconsequential changes among the weakly correlated PPR comprising the vast majority of cases. In the moderately and strongly correlated sectors, there are a handful of more substantive $|\rho|$ changes worth mention, but none of which warrant changes to our working classification scheme. In most of these cases, we observed an increase in $|\rho(x,y)|$ when this quantity was computed using the equilibrium structures only—this is an expected change as the omission of the non-equilibrium structures *tends* to increase cov(x,y) while simultaneously decreasing both σ_x and σ_y (*cf.* eqn (1)); see Fig. S1† and surrounding discussion for more details. Interestingly, the largest difference was observed for (E_{LUMO}, E_{gap}), for which $|\rho|$ unexpectedly *decreased* from 0.84 to 0.67 when computed using the equilibrium structures only. This change in $|\rho|$ highlights the non-trivial influence that non-equilibrium molecular structures can have on the observed pairwise correlations in molecular property space, and is (somewhat) easier to rationalize by considering it as an unexpected *increase* when the non-equilibrium structures were included in the evaluation of $|\rho|$. In this case, the non-equilibrium structures lead to a disproportionate increase in cov(E_{LUMO}, E_{gap}) (*i.e.*, a measure of the diagonal spread) relative to the simultaneous increase in $\sigma_{E_{LUMO}}\sigma_{E_{gap}}$ (*i.e.*, a measure of the axis-aligned spreads), which results in an overall increase in $|\rho|$ but no change to the classification of this PPR (see Fig. S1†).

Although the inclusion of non-equilibrium molecular structures did not meaningfully alter the scheme used to classify the PPR in this work, it is important to recognize that their (partial or full) inclusion can lead to non-trivial effects on observed property values (and hence the rational design of molecules with targeted properties). To quantify this effect, we critically assessed the coefficient of variation (or relative standard deviation) $c_v \equiv \sigma_x/\bar{x}$, *i.e.*, the ratio of the standard deviation σ_x to the mean value $\bar{x}$ for a given property x, for several representative extensive (*e.g.*, E_{AT}, E_{MBD}, α) and intensive (*e.g.*, E_{gap}, μ, E_{HOMO}) properties (see Fig. S2†). In doing so, we found that the extensive properties have $c_v \ll 0.1$, and are therefore essentially unaffected by the structural diversity in the QM7-X non-equilibrium conformations, while the intensive properties are much more sensitive to such structural variations. Both of these findings are consistent with the fact that the non-equilibrium structural variations in QM7-X are largely perturbative (with respect to the corresponding equilibrium structures), and are characterized by changes to the molecular geometry that conserve atom connectivity and leave the molecule intact (*i.e.*, non-equilibrium bond lengths, bond angles, and dihedrals). Hence, this analysis provides strong support for including such semi-local molecular PES information when characterizing and exploring the range/distribution of property values (particularly for intensive properties) that are accessible by a given molecule.

With only a handful of exceptions, the above analysis of the pairwise PPR in the sector of molecular property space spanned by the small (primarily organic) molecules in QM7-X demonstrates that most QM properties are effectively uncorrelated. While one might initially view this as a challenge for rational molecular design, we would argue that this finding highlights an intrinsic flexibility—or "freedom of design"—that exists in CCS, wherein there seems to be very few limitations preventing a molecule from simultaneously exhibiting any pair of properties considered in Fig. 1. Since this "freedom of design" conjecture has profound implications regarding the existence and uniqueness of molecules with a diverse array of targeted properties, it will be critically analyzed and assessed throughout the remainder of this work.

3.2 Structural and compositional dependence of molecular property space

While the complex set of pairwise PPR analyzed in Sec. 3.1 suggests that one has "freedom of design" when searching for small molecules with a targeted pair of properties, exactly how this intrinsic degree of flexibility is related to the molecular structure and chemical composition in this sector of CCS—the tunable knobs in the molecular design process—still requires further investigation. To do so, we start by considering a pair of properties that exhibited a moderate-to-strong degree of correlation in our previous analysis (*i.e.*, (E_{MBD}, E_{AT}) with $|\rho| = 0.87$), as this relatively narrow sector of molecular property space is expected to have less intrinsic flexibility and will therefore serve as a more challenging test of our "freedom of design" conjecture. More specifically, we consider thermally-averaged versions of these quantities (*i.e.*, ($\langle E_{MBD}\rangle$, $\langle E_{AT}\rangle$) at T = 300 K) as thermally-averaged properties are commonly employed during molecular design procedures. In this case, thermal averaging increases $|\rho|$ from 0.87 to 0.92, making these properties strongly correlated according to our classification system. Here, we note that the thermally-averaged $|\rho|$ for these properties is effectively identical to that computed using the equilibrium structures only (see Fig. S1(d)†), which is not surprising as thermally-averaged property values (at T = 300 K) tend to be very similar to their corresponding equilibrium values for the majority of molecules and properties in QM7-X (see Fig. S2†).

As depicted in Fig. 2(a), the range of $|\langle E_{MBD}\rangle|$ and $\langle E_{AT}\rangle$ values (0.02–0.48 eV and 19.3–103.3 eV, respectively) is quite large, indicating that the molecules in QM7-X are quite diverse and cover a sizeable sector of CCS. The molecules with the lowest $|\langle E_{MBD}\rangle|$ and $\langle E_{AT}\rangle$ values are small hydrocarbons such as CH_4 (∼0.02 eV and ∼19.3 eV) and C_2H_2 (∼0.02 eV and ∼19.9 eV),

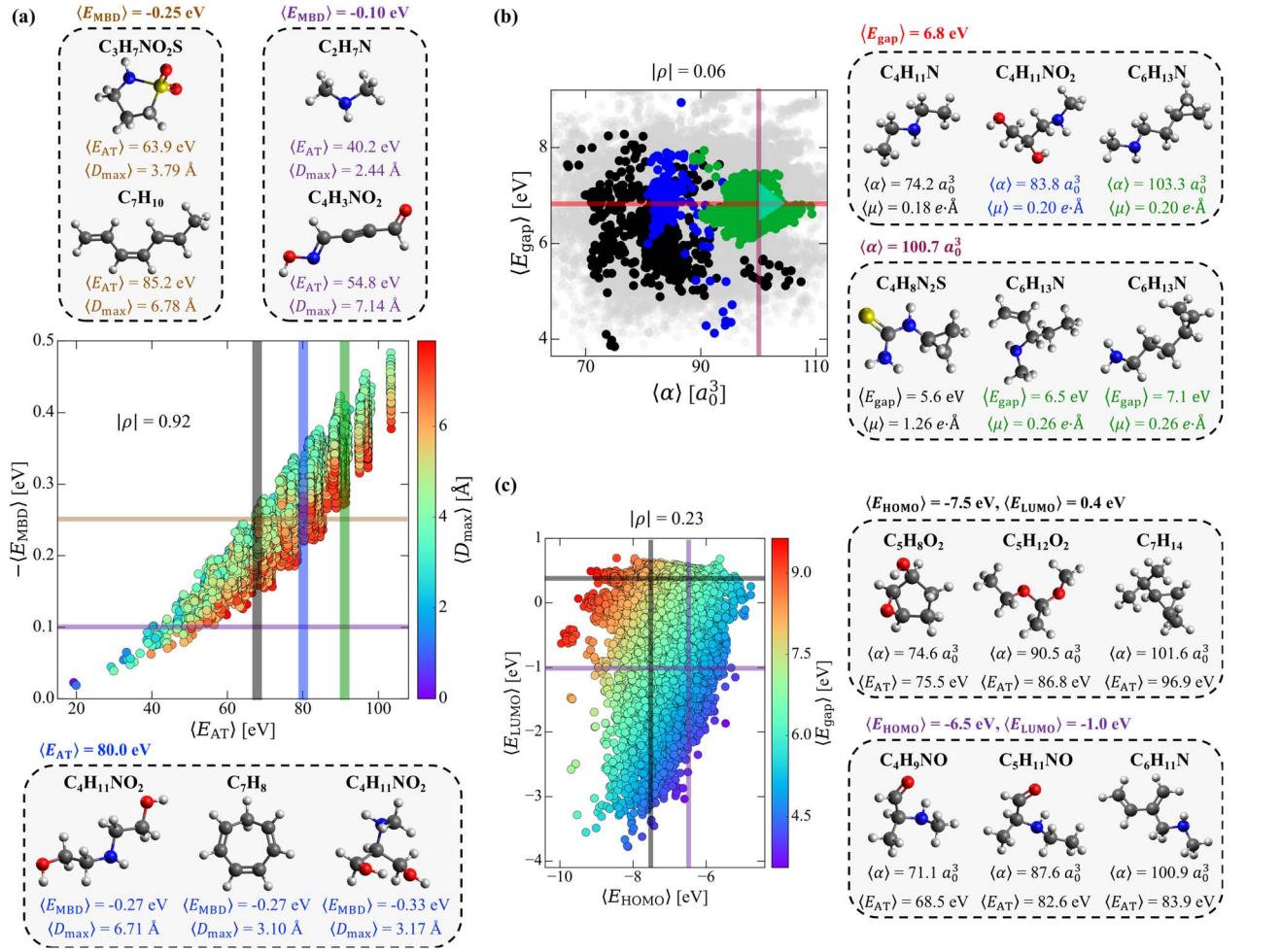

Fig. 2 Structural and compositional dependence of molecular property space. (a) Correlation plot between the thermally-averaged ($T = 300$ K) MBD vdW/dispersion energy ($-\langle E_{MBD}\rangle$) and atomization energy ($\langle E_{AT}\rangle$) for the equilibrium structures in QM7-X, with each point colored according to the corresponding thermally-averaged maximum distance between heavy/non-hydrogen atoms ($\langle D_{max}\rangle$); see Sec. 2 for more details. Also depicted are select molecules from the $\langle E_{MBD}\rangle = -0.25 \pm 0.01$ eV and $\langle E_{MBD}\rangle = -0.10 \pm 0.01$ eV windows (*top inset*) and $\langle E_{AT}\rangle = 80 \pm 0.2$ eV window (*bottom inset*). (b) Correlation plot between the thermally-averaged ($T = 300$ K) HOMO–LUMO gap ($\langle E_{gap}\rangle$) and isotropic molecular polarizability ($\langle\alpha\rangle$), with each point colored according to the $\langle E_{AT}\rangle$ windows in (a). Gray points in the background correspond to values outside these $\langle E_{AT}\rangle$ windows. Also depicted are select molecules from the $\langle E_{gap}\rangle = 6.8 \pm 0.02$ eV and $\langle\alpha\rangle = 100.7 \pm 0.3\ a_0^3$ windows (*right insets*); a cyan triangle highlights the location of the three $C_6H_{13}N$ isomers in the correlation plot. (c) Correlation plot between the thermally-averaged ($T = 300$ K) HOMO energy ($\langle E_{HOMO}\rangle$) and LUMO energy ($\langle E_{LUMO}\rangle$), with each point colored according to ($\langle E_{gap}\rangle$). Also depicted are select molecules from the $\langle E_{HOMO}\rangle = -7.5 \pm 0.04$ eV, $\langle E_{LUMO}\rangle = 0.4 \pm 0.04$ eV and $\langle E_{HOMO}\rangle = -6.5 \pm 0.04$ eV, $\langle E_{LUMO}\rangle = -1.0 \pm 0.04$ eV windows (*right inset*). This detailed analysis highlight numerous instances where two structurally and/or compositionally distinct molecules share multiple (*i.e.*, 2–4) extensive and/or intensive properties, thereby providing even more compelling evidence that "freedom of design" is a fundamental property of CCS.

while the largest values correspond to C_7H_{16} isomers/conformers (~0.48 eV and ~103.3 eV); the molecules in QM7-X containing second-row elements (*i.e.*, S and Cl) tend to have intermediate values for these quantities (see Fig. S3†). Despite the strong correlation between these extensive properties, there is still visible dispersion in Fig. 2(a), which indicates that diverse ($\langle E_{MBD}\rangle$, $\langle E_{AT}\rangle$) combinations are possible, *i.e.*, for a fixed value of one property, there is considerable flexibility in the value of the other. From this plot, one can also see that this dispersion is fairly well-correlated with $\langle D_{max}\rangle$, a quantity that measures the spatial extent of each molecule *via* the thermally-averaged *maximum* pairwise distance between heavy/non-hydrogen atoms in a given molecular geometry (see Sec. 2). To explore these points further, we characterized the structure and composition of the molecules contained in two fixed $\langle E_{MBD}\rangle$ windows, $\langle E_{MBD}\rangle = -0.25 \pm 0.01$ eV and $\langle E_{MBD}\rangle = -0.10 \pm 0.01$ eV, which represent the 50th and 20th percentiles of the observed vdW/dispersion energy spectrum. In doing so, we were able to easily find molecules with markedly distinct structures (*i.e.*, compact *vs.* extended as quantified by $\langle D_{max}\rangle$) and chemical compositions with the same $\langle E_{MBD}\rangle$ but completely different $\langle E_{AT}\rangle$. This is another manifestation of "freedom of design" in CCS, and is clearly illustrated by the $C_3H_7NO_2S$ and C_7H_{10} isomers in the top inset of Fig. 2(a): while both are located

in the $\langle E_{MBD}\rangle = -0.25 \pm 0.01$ eV window (at opposite edges of the data dispersion), their $\langle E_{AT}\rangle$ values differ by more than 20 eV. Since $|\langle E_{MBD}\rangle|$ is an extensive property that tends to increase with the number of atoms in a molecule and decrease with molecular volume/spatial extent, $C_3H_7NO_2S$ (a compact molecule with less atoms, $\langle D_{max}\rangle = 3.79$ Å) and C_7H_{10} (an extended molecule with more atoms, $\langle D_{max}\rangle = 6.78$ Å) represent a non-trivial compromise between these two effects that results in similar $\langle E_{MBD}\rangle$. In the same breath, the sizeable difference in $\langle E_{AT}\rangle$ between these molecules can be primarily attributed to the larger number of atoms in C_7H_{10} as well as its conjugated/extended π-system, which further stabilizes this hydrocarbon and increases $\langle E_{AT}\rangle$. When analyzing the less dense $\langle E_{MBD}\rangle = -0.10 \pm 0.01$ eV window, one can just as easily find another distinct pair of molecules (again located at the edges of the data dispersion) that exhibit markedly different $\langle E_{AT}\rangle$. Here, we observe an ≈15 eV $\langle E_{AT}\rangle$ difference between C_2H_7N and $C_4H_3NO_2$ (see Fig. 2(a), top inset), which can be rationalized by the larger number of heavy atoms and more complex bonding motifs (*e.g.*, C=O, C=N, C≡C) in $C_4H_3NO_2$.

With such dispersion in the ($\langle E_{MBD}\rangle$, $\langle E_{AT}\rangle$) correlation plot, a similar degree of flexibility also exists when holding $\langle E_{AT}\rangle$ fixed. For instance, analyzing the molecules within the $\langle E_{AT}\rangle = 80 \pm 0.2$ eV window (a region of high density in this correlation plot, see Fig. S3†) uncovered a group of molecules with different structures and/or chemical compositions and a range of $\langle E_{MBD}\rangle$ (*e.g.*, the $C_4H_{11}NO_2$ and C_7H_8 isomers in the bottom inset of Fig. 2(a)). When comparing the extended ($\langle D_{max}\rangle = 6.71$ Å) and compact ($\langle D_{max}\rangle = 3.17$ Å) $C_4H_{11}NO_2$ isomers, the latter exhibits a more negative $\langle E_{MBD}\rangle$; this is consistent with the more sizeable vdW/dispersion energy contributions that arise from the relatively closer non-bonded atoms in compact molecular geometries. In contrast, the extended $C_4H_{11}NO_2$ isomer has the same $\langle E_{MBD}\rangle$ (and $\langle E_{AT}\rangle$) as the more compact ring-like C_7H_8 hydrocarbon ($\langle D_{max}\rangle = 3.10$ Å)—another illustrative example of the non-trivial compromises made between the number of atoms, chemical composition, and volume/spatial extent of a molecule in determining $\langle E_{MBD}\rangle$. This example also illustrates another important aspect of "freedom of design" in CCS, *i.e.*, that two distinct molecules can share multiple physicochemical properties (*vide infra*). Interestingly, despite having very similar $\langle D_{max}\rangle$, the compact $C_4H_{11}NO_2$ isomer has a more negative $\langle E_{MBD}\rangle$ than the compact ring-like C_7H_8 isomer—a result of more nuanced topological effects (*i.e.*, packed/globular *vs.* void space) on the vdW/dispersion interactions in molecules.[78]

By considering $\langle D_{max}\rangle$ in this analysis, we have partially accounted for the fact that $\langle E_{MBD}\rangle$ and $\langle E_{AT}\rangle$ are extensive properties. However, one can perform the same analysis after explicitly normalizing these properties (*i.e.*, $\langle E_{MBD}\rangle \rightarrow \langle E'_{MBD}\rangle$ and $\langle E_{AT}\rangle \rightarrow \langle E'_{AT}\rangle$) with respect to different size-dependent quantities (see Sec. 2). As depicted in Fig. S4†, the degree of correlation between $\langle E'_{MBD}\rangle$ and $\langle E'_{AT}\rangle$ *strongly* depends on the chosen normalization quantity and can therefore be quite variable, ranging from a slight increase in $|\rho|$ (*i.e.*, $|\rho| = 0.92 \rightarrow 0.93$, when normalized with respect to the thermally-averaged molecular volume $\langle V\rangle$) to a substantial decrease in $|\rho|$ (*i.e.*, $|\rho| = 0.92 \rightarrow 0.37$, when normalized with respect to the total number of atoms N_{atoms}). Despite such sizeable changes to the degree of correlation between these properties, we were still able to find example molecules located within fixed $\langle E'_{MBD}\rangle$ and $\langle E'_{AT}\rangle$ windows (for either normalization protocol) that were analogous to those obtained using the extensive variants (*cf.* the top and bottom insets in Fig. S4(a), (b)† and 2(a)). In other words, we were able to find structurally and/or compositionally distinct molecules with the same $\langle E'_{MBD}\rangle$ but different $\langle E'_{AT}\rangle$ (and *vice versa*), as well as markedly distinct molecules sharing two (or more) normalized properties (see Fig. S4† and surrounding discussion). As such, these findings provide strong evidence that our "freedom of design" conjecture is quite robust and effectively independent of the use and choice of normalization protocol when dealing with extensive properties. For simplicity, we will therefore continue our analysis using non-normalized extensive properties for the remainder of this work.

Having examined the intrinsic flexibility existing between two extensive properties, we now turn our attention to the 2D projection of molecular property space corresponding to $\langle E_{gap}\rangle$ (an intensive property) and $\langle \alpha\rangle$ (an extensive property). Fig. 2(b) depicts the corresponding ($\langle E_{gap}\rangle$, $\langle \alpha\rangle$) correlation plot (with points colored according to the $\langle E_{AT}\rangle$ windows in Fig. 2(a)); with a particularly low correlation coefficient of $|\rho| = 0.06$, these properties are effectively uncorrelated. Interestingly, we still find a large range of $\langle E_{gap}\rangle$ and $\langle \alpha\rangle$ (*i.e.*, 3.8–8.4 eV and 68.0–110.0 a_0^3), even though we are only considering the molecules contained in three narrow $\langle E_{AT}\rangle$ windows. Similar to the ($\langle E_{MBD}\rangle$, $\langle E_{AT}\rangle$) analysis performed above, we will fix one property and consider the flexibility in the other (and *vice versa*). Starting with the $\langle E_{gap}\rangle = 6.8 \pm 0.02$ eV window (an intermediate HOMO–LUMO gap in this dataset), we again find molecules with distinct structures and compositions (see right inset in Fig. 2(b)) exhibiting a wide range of $\langle \alpha\rangle$ (*i.e.*, 74.2 a_0^3 ($C_4H_{11}N$) to 103.3 a_0^3 ($C_6H_{13}N$)) and $\langle E_{AT}\rangle$ (*i.e.*, 68.4 eV ($C_4H_{11}N$) to 90.0 eV ($C_6H_{13}N$)). Interestingly, the depicted $C_4H_{11}NO_2$ and $C_6H_{13}N$ isomers also share the same $\langle \mu\rangle$ (in addition to $\langle E_{gap}\rangle$), which is another example of the flexibility one has when searching for distinct molecules that share multiple physicochemical properties (see Sec. 3.3). Turning now to the fixed $\langle \alpha\rangle = 100.7 \pm 0.3$ a_0^3 window, we similarly found a set of distinct molecules (see right inset in Fig. 2(b)) that exhibit a wide range of $\langle E_{gap}\rangle$ (*i.e.*, 5.6 eV ($C_4H_8N_2S$) to 7.1 eV ($C_6H_{13}N$)) and $\langle E_{AT}\rangle$ (*i.e.*, 68.3 eV ($C_4H_8N_2S$) to 90.2 eV ($C_6H_{13}N$)). Within this group of molecules, we also found a two-molecule set (comprised of the two $C_6H_{13}N$ isomers) that share four (extensive and intensive) properties: $\langle E_{AT}\rangle$, $\langle E_{MBD}\rangle$, $\langle \alpha\rangle$, and $\langle \mu\rangle$; in Sec. 3.3, we will show that this is not a rare occurrence or "cherry-picked example", as there are thousands of *three*- and *four*-molecule sets (among the ≈41k equilibrium molecules in QM7-X) which share four properties. When considered together, the *three* $C_6H_{13}N$ isomers depicted in the right insets of Fig. 2(b) also provide a simple but illustrative example of the inherent flexibility that exists throughout CCS. Through small and directed changes in the molecular structure—just one of the tunable knobs in the molecular design process—these constitutional isomers allow one to traverse a path through ($\langle E_{gap}\rangle$, $\langle \alpha\rangle$)-space (*i.e.*, the cyan triangle in Fig. 2(b)) in which an increase in $\langle E_{gap}\rangle$ can be accompanied by an increase, a decrease, or no change in $\langle \alpha\rangle$.

As a final case study, we now consider the flexibility one has in finding distinct molecules with the same E_{HOMO} and E_{LUMO}—two fundamentally important intensive properties that govern chemical reactivity and electron transfer in molecules. This choice of intensive properties poses an additional and more nuanced challenge to our "freedom of design" conjecture, since the number of structurally and/or compositionally distinct molecules sharing two properties ($\langle E_{HOMO}\rangle$ and $\langle E_{LUMO}\rangle$) will be significantly less than those sharing the same $\langle E_{gap}\rangle$ (*vide supra*). To proceed, we chose two distinct points in the ($\langle E_{HOMO}\rangle$, $\langle E_{LUMO}\rangle$)-sector of molecular property space (see Fig. 2(c)). The first point corresponds to the window delineated by $\langle E_{HOMO}\rangle = -7.5 \pm 0.04$ eV and $\langle E_{LUMO}\rangle = 0.4 \pm 0.04$ eV, where we were able to find a set of structurally and compositionally distinct molecules, including (but not limited to): $C_5H_8O_2$ (a cyclic molecule containing both epoxide and hydroxyl functional groups), $C_5H_{12}O_2$ (an asymmetric ketal), and C_7H_{14} (a cyclopropyl-containing alkane) that share these $\langle E_{HOMO}\rangle$ and $\langle E_{LUMO}\rangle$ values (see top right inset of Fig. 2(c)). With $\langle E_{AT}\rangle$ and $\langle \alpha\rangle$ ranging from 75.5–96.9 eV and 74.6–101.6 a_0^3, this subset of molecules is rather diverse and indicative of considerable flexibility in this constrained sector of molecular property space. At the second of these points ($\langle E_{HOMO}\rangle = -6.5 \pm 0.04$ eV and $\langle E_{LUMO}\rangle = -1.0 \pm 0.04$ eV), we were again able to find a diverse set of molecules with a similarly wide range of $\langle E_{AT}\rangle$ (68.5–83.9 eV) and $\langle \alpha\rangle$ (71.1–100.9 a_0^3). In this case, some of the molecules that share these $\langle E_{HOMO}\rangle$ and $\langle E_{LUMO}\rangle$ values include: C_4H_9NO (an amine-containing aldehyde), $C_5H_{11}NO$ (another amine-containing aldehyde), and $C_6H_{11}N$ (an amine-containing conjugated alkene); see bottom right inset of Fig. 2(c). In many ways, this subset of molecules also illustrates how small and directed changes underlie rational molecular design. For one, the simple addition of a methyl group which converts C_4H_9NO to $C_5H_{11}NO$ can be used to change the extensive properties (*e.g.*, $\langle \alpha\rangle$ and $\langle E_{AT}\rangle$) without modifying the intensive properties (*e.g.*, $\langle E_{HOMO}\rangle$, $\langle E_{LUMO}\rangle$, $\langle E_{gap}\rangle$). In the same breath, more complex changes (*i.e.*, functional group modifications, alchemical changes, increases in conjugation) can be used to induce selective and non-trivial modifications to some extensive properties (*e.g.*, $\langle \alpha\rangle$) while leaving others (*e.g.*, $\langle E_{AT}\rangle$) effectively unchanged.

In summary, these three complementary case studies show that the "freedom of design" conjecture proposed herein applies rather generally to both extensive and intensive properties, independent of whether these properties exhibit strong or weak mutual correlations. While also demonstrating that this conjecture is quite robust and effectively independent of the use and choice of normalization protocol when considering extensive properties, our analysis also indicates that the normalization protocol can potentially be used to enhance (or perhaps optimize) the flexibility in CCS when searching for distinct molecules with a targeted set of properties.

3.3 Multi-property analysis: exploring more complex manifolds of molecular property space

The fact that we were able to find numerous examples of distinct molecules that share two (or more) extensive and/or intensive properties in restricted sectors (*i.e.*, selected windows) of QM7-X molecular property space in Sec. 3.2 provides strong support for an intrinsic flexibility in CCS that can be leveraged when searching for molecules with targeted QM properties. In this section, we demonstrate that these are not rare occurrences (or "cherry-picked examples") by lifting such restrictions and exhaustively enumerating the number of N-molecule sets (*i.e.*, sets containing $N = 2$, 3, 4 unique molecules) that share two, three, or four of the following thermally-averaged ($T = 300$ K) extensive ($P_1 = \langle E_{AT}\rangle$ and $P_2 = \langle \alpha\rangle$) and/or intensive ($P_3 = \langle E_{gap}\rangle$ and $P_4 = \langle \mu\rangle$) properties. This

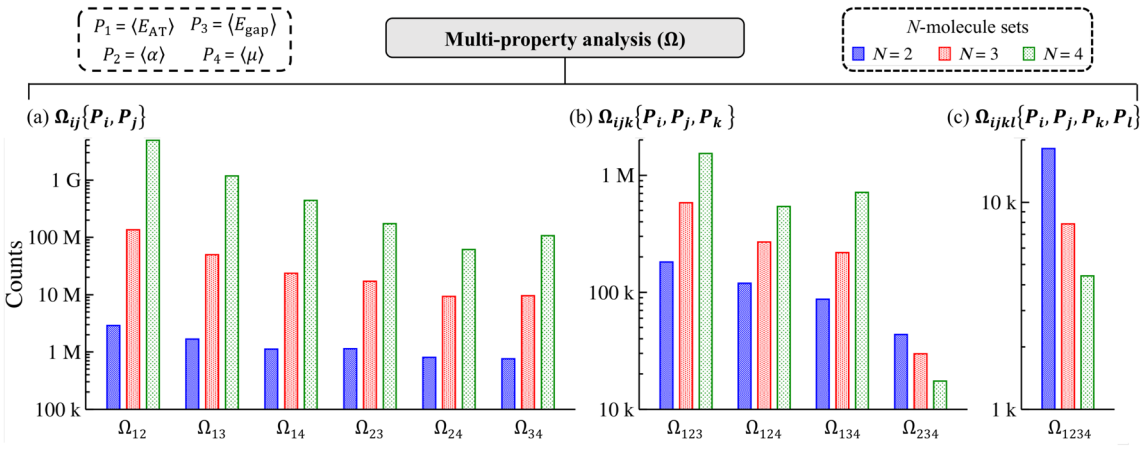

Fig. 3 Multi-property analysis in molecular property space. Three different multi-property analyses (Ω) were performed to exhaustively enumerate the number of N-molecule sets (*i.e.*, sets containing $N = 2$, 3, 4 unique molecules taken from the ≈41k equilibrium molecules in QM7-X) that share: (a) two properties ($\Omega_{ij}\{P_i, P_j\}$), (b) three properties ($\Omega_{ijk}\{P_i, P_j, P_k\}$), and (c) four properties ($\Omega_{ijkl}\{P_i, P_j, P_k, P_l\}$). In these analyses, we considered the following thermally-averaged extensive ($P_1 = \langle E_{AT}\rangle$ and $P_2 = \langle \alpha\rangle$) and intensive ($P_3 = \langle E_{gap}\rangle$ and $P_4 = \langle \mu\rangle$) properties at $T = 300$ K. For instance, three-property analysis at the $\Omega_{134}\{\langle E_{AT}\rangle, \langle E_{gap}\rangle, \langle \mu\rangle\}$ (or Ω_{134}) level would involve enumerating all 2-, 3-, and 4-molecule sets that share the same $\langle E_{AT}\rangle$, $\langle E_{gap}\rangle$, and $\langle \mu\rangle$ values. The remarkably large number of N-molecule sets found during these multi-property analyses provide direct and compelling evidence for "freedom of design" in CCS, in which one has a substantial degree of flexibility when searching for structurally and/or compositionally diverse molecules that share multiple physicochemical properties.

enumeration will be performed on the ≈41k equilibrium molecules in QM7-X and property-specific thresholds will be used to identify each *N*-molecule set (see Sec. 2 for more details).

We start this analysis at the two-property tier, which consists of six different two-property combinations/levels (denoted by $\Omega_{ij}\{P_i, P_j\}$ throughout). Even from a quick glance at Fig. 3(a), one can see that the number of *N*-molecule sets that share any pair of these four properties is quite large, with total counts ranging from just under 1M to well over 4G. By far, the largest counts resulted from enumeration at the $\Omega_{12}\{\langle E_{AT}\rangle, \langle\alpha\rangle\}$ level, which yielded ≈2.9M unique 2-molecule sets, ≈134.5M unique 3-molecule sets, and ≈4.9G unique 4-molecule sets. In other words, there are nearly three million unique pairs of molecules (among the ≈863M possible unique molecular pairs) that have the same $\langle E_{AT}\rangle$ and $\langle\alpha\rangle$. In general, these counts tend to decrease as the extensive properties are replaced by the intensive properties, with the smallest counts resulting from enumeration at the $\Omega_{24}\{\langle\alpha\rangle, \langle\mu\rangle\}$ and $\Omega_{34}\{\langle E_{gap}\rangle, \langle\mu\rangle\}$ levels (although there are still millions of 2-, 3-, and 4-molecule sets in both cases). We also found that the number of *N*-molecule sets consistently increased with *N* across the entire two-property tier, such that the number of 4-molecule sets > the number of 3-molecule sets > the number of 2-molecule sets for all two-property combinations. Both of these trends illustrate the remarkable degree of flexibility that one has when searching for distinct molecules sharing any two of these properties, and will be discussed in more detail below.

While the number of *N*-molecule sets that share three properties are considerably less than those sharing two, the total counts at the three-property tier are still quite large and range from 10k to 1M (see Fig. 3(b)). The largest counts of 2-, 3-, and 4-molecule sets were found at the $\Omega_{123}\{\langle E_{AT}\rangle, \langle\alpha\rangle, \langle E_{gap}\rangle\}$ level (which contains two extensive and one intensive properties), while the smallest counts were found at the $\Omega_{234}\{\langle\alpha\rangle, \langle E_{gap}\rangle, \langle\mu\rangle\}$ level (which contains one extensive and two intensive properties). While we still observe a consistent rise in the total count of *N*-molecule sets from $N = 2$ to $N = 4$ at the $\Omega_{123}\{\langle E_{AT}\rangle, \langle\alpha\rangle, \langle E_{gap}\rangle\}$, $\Omega_{124}\{\langle E_{AT}\rangle, \langle\alpha\rangle, \langle\mu\rangle\}$, and $\Omega_{134}\{\langle E_{AT}\rangle, \langle E_{gap}\rangle, \langle\mu\rangle\}$ levels, this trend is completely reversed at the $\Omega_{234}\{\langle\alpha\rangle, \langle E_{gap}\rangle, \langle\mu\rangle\}$ level. This observation can be explained by considering the combinatorics that determine the number of *N*-molecule sets that share a given array of properties. Although our analysis here focuses on *N*-molecule sets with $N = 2, 3, 4$, enumeration of the ≈41k equilibrium molecules in QM7-X often finds significantly larger clusters (or *parent sets*) that contain $M > N$ molecules sharing a given array of properties. Since the number of *N*-molecule sets that can be formed from one of these *M*-molecule clusters is given by $C(M,N) = \frac{M!}{N!(M-N)!}$, each cluster with $M \geq 8$ will generate more 4-molecule sets than 3-molecule sets and more 3-molecule sets than 2-molecule sets. Hence, the observed increase in the total count of *N*-molecule sets from $N = 2$ to $N = 4$ will start to break down when $M < 8$, which indicates that enumeration at the $\Omega_{234}\{\langle\alpha\rangle, \langle E_{gap}\rangle, \langle\mu\rangle\}$ level tends to find *M*-molecule clusters containing fewer molecules than enumeration at the three other $\Omega_{ijk}\{P_i, P_j, P_k\}$ levels. The most likely explanation for these observed trends is twofold: (i) since the ≈41k equilibrium molecules in QM7-X correspond to ≈7k different molecular formulae, structural changes are sampled to a larger extent than compositional changes in the molecules enumerated in this analysis (see Sec. 2 and ref. 34); and (ii) intensive properties tend to be more sensitive to such structural changes than extensive properties.

Quite interestingly, the number of *N*-molecule sets that share all four of these properties are also quite large. As depicted in Fig. 3(c), enumeration at the $\Omega_{1234}\{\langle E_{AT}\rangle, \langle\alpha\rangle, \langle E_{gap}\rangle, \langle\mu\rangle\}$ level—the most complex manifold of molecular property space considered in this work—found ≈20k 2-molecule sets, ≈8k 3-molecule sets, and ≈4k 4-molecule sets. Intrigued by these results, we also considered the same analysis with $\langle E_{HOMO}\rangle$ and $\langle E_{LUMO}\rangle$ (instead of $\langle E_{gap}\rangle$ and $\langle\mu\rangle$) as the intensive properties (see Fig. S5†). In doing so, we found that the number of *N*-molecule sets at the three- and four-property tiers in this more stringent case are actually larger than those shown in Fig. 3(b) and (c). For instance, there are nearly 2.5× more 4-molecule sets (*i.e.*, ≈10k) at the $\Omega_{1234}\{\langle E_{AT}\rangle, \langle\alpha\rangle, \langle E_{HOMO}\rangle, \langle E_{LUMO}\rangle\}$ level than the $\Omega_{1234}\{\langle E_{AT}\rangle, \langle\alpha\rangle, \langle E_{gap}\rangle, \langle\mu\rangle\}$ level.

In summary, this tiered multi-property analysis has demonstrated that it is quite feasible to find structurally and/or compositionally distinct molecules that share 2–4 extensive/intensive ground-state and response properties—yet another manifestation that "freedom of design" is a fundamental and emergent property of CCS. By considering the decay rate in the number of *N*-molecule sets, we also estimate that 7–10 properties are needed to *uniquely* identify each molecule in the sector of CCS spanned by QM7-X. This quantity corresponds to the effective dimensionality of a small (primarily organic) molecule in molecular property space and can potentially be used to guide the length of property-based molecular features (or fingerprints) for use in ML applications.

3.4 Multi-property optimization: finding optimal Pareto fronts in molecular property space

In the molecular design process, which often involves the simultaneous optimization of multiple (*i.e.*, typically two or more) physicochemical properties, Pareto fronts (or frontiers) represent the so-called Pareto-optimal solutions for which no single property can be improved without degrading the other(s). Pareto fronts have therefore been used in a number of different fields (*e.g.*, economics, medicine, materials science, chemical engineering)[79–81] and have given rise to evolutionary multi-objective optimization.[48,49] In this work, we use this approach in the complex manifold of molecular property space defined by $\langle E_{AT}\rangle$, $\langle\alpha\rangle$, and $\langle E_{gap}\rangle$ to search for molecules in QM7-X with simultaneously large $\langle\alpha\rangle$ and $\langle E_{gap}\rangle$ values. Here, we note in passing 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 $\langle\mu\rangle$).

To begin, we partitioned the QM7-X molecules in the weakly correlated ($\langle\alpha\rangle$, $\langle E_{gap}\rangle$)-space according to the following $\langle E_{AT}\rangle$

windows: [40,50) eV, [60,70) eV, and [70,80) eV (see Fig. 4); these windows encompass the molecular diversity in QM7-X (see Fig. S3†), and contain molecules with distinct chemical compositions and a wide range of $\langle\alpha\rangle$ and $\langle E_{gap}\rangle$ values. The Pareto fronts resulting from independent multi-property optimizations carried out in each of these $\langle E_{AT}\rangle$ windows are plotted in Fig. 4. Each of these fronts represents a unique and non-trivial path through CCS consisting of a sequence of molecules connected by structural and/or compositional changes that define the optimal balance between these properties. While consecutive molecules on a given front can often be rationalized *a posteriori* using chemical/physical intuition, several interesting and unexpected molecules appear in this analysis, reflecting the freedom one has when designing molecules with a targeted array of properties.

Consider the Ⓐ → Ⓑ front, which contains 11 diverse molecular structures with varied compositions and $\langle E_{AT}\rangle \in [40,50)$ eV. This front starts with three constitutional isomers of $C_4H_2N_2S$, each of which contains a terminal alkyne (ethynyl group) directly adjacent to an aromatic thiadiazole ring. Such conjugation facilitates charge delocalization and π-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 $\langle E_{gap}\rangle$. However, $\langle E_{gap}\rangle$ is more sensitive to the structural differences connecting these isomers and can be tuned by ≈0.6 eV by changing the relative positions of the heteroatoms in the thiadiazole ring. Continuing along this front, a linear and highly conjugated molecule (penta-2,4-diynenitrile, C_5HN) appears, with a structure and composition that is quite different from the $C_4H_2N_2S$ isomers. Despite such differences, this molecule has $\langle\alpha\rangle$ and $\langle E_{gap}\rangle$ values that are both very similar to its Pareto-adjacent $C_4H_2N_2S$ isomers, which again illustrates the flexibility provided by CCS when searching for distinct molecules sharing multiple properties. Ethynyl sulfone ($C_4H_2O_2S$) is the next molecule on this front, and contains two terminal alkynes connected *via* a central sulfonyl (SO_2) moiety in a kinked arrangement—another large change in both structure and composition when compared to C_5HN. This non-linear molecular geometry reduces π-electron mobility and charge delocalization, resulting in a rather large change (*i.e.*, >1 eV) in $\langle E_{gap}\rangle$. In the same breath, this molecule shares a nearly identical $\langle\alpha\rangle$ with C_5HN—a simple and illustrative example of "freedom of design" provided by this non-trivial path through CCS. Also interesting is the natural emergence of the rational molecular design process when *retrospectively* analyzing a Pareto front. To see this, consider the sequence of small directed changes needed to arrive at the next three molecules on the Ⓐ → Ⓑ front. Since the polarizability of a N atom is smaller than that of

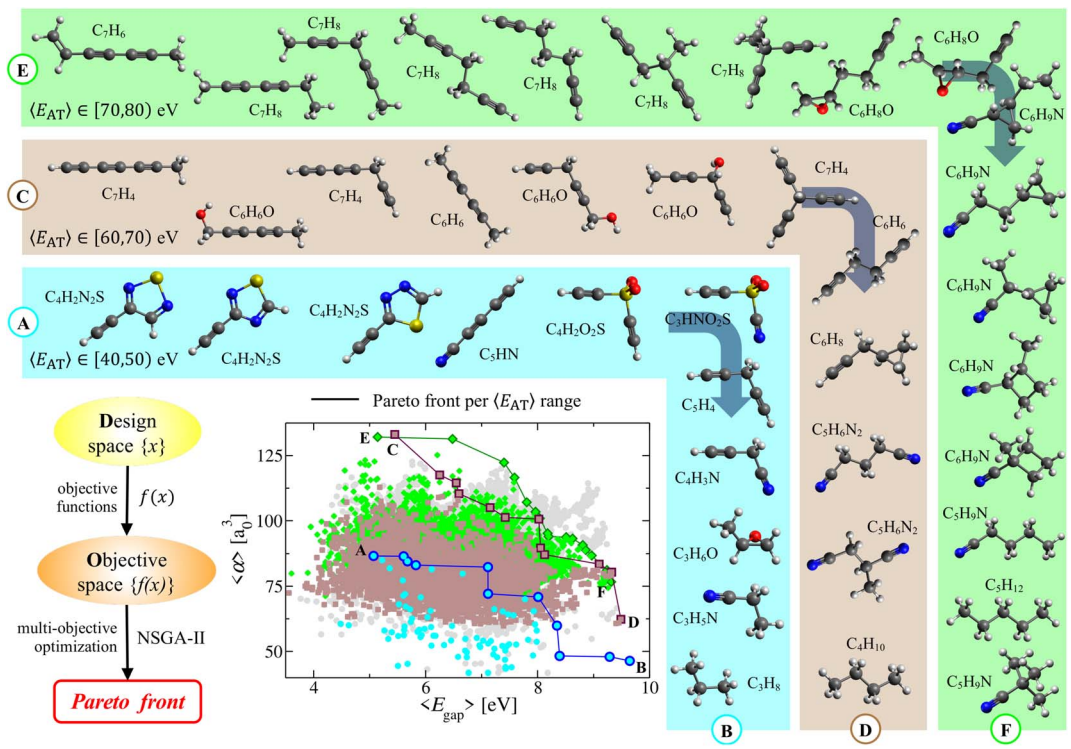

Fig. 4 Multi-property optimization: Pareto front analysis in molecular property space. Pareto (multi-property) optimization was performed on the thermally-averaged ($\langle E_{AT}\rangle$, $\langle E_{gap}\rangle$, $\langle\alpha\rangle$)-manifold of molecular property space ($T = 300$ K) to identify molecules with simultaneously optimal $\langle\alpha\rangle$ and $\langle E_{gap}\rangle$ values. Depicted is a correlation plot between $\langle\alpha\rangle$ and $\langle E_{gap}\rangle$, with each point colored according to the corresponding $\langle E_{AT}\rangle$ range (cyan: $\langle E_{AT}\rangle \in [40,50)$ eV; tan: $\langle E_{AT}\rangle \in [60,70)$ eV; light green: $\langle E_{AT}\rangle \in [70,80)$ eV); gray points in the background correspond to values outside these $\langle E_{AT}\rangle$ windows. Optimal Pareto fronts corresponding to each of these $\langle E_{AT}\rangle$ windows are provided as highlighted points (connected *via* solid lines) in the correlation plot; see Sec. 2 for a more detailed description of the optimization procedure. Also depicted are the QM7-X molecules located along each of these Pareto fronts (colored insets), which correspond to non-trivial paths through CCS defined by small directed structural/compositional changes that yield simultaneously optimal values for these properties.

C–H, replacing an ethynyl (C≡C–H) with a nitrile (C≡N) group (*i.e.*, $C_4H_2O_2S \rightarrow C_3HNO_2S$) can be used to decrease $\langle\alpha\rangle$ while leaving $\langle E_{gap}\rangle$ effectively unchanged. Alternatively, one could instead replace SO_2 with a more insulating methylene (CH_2) group (*i.e.*, $C_4H_2O_2S \rightarrow C_5H_4$) to decrease $\langle\alpha\rangle$ while simultaneously increasing $\langle E_{gap}\rangle$ by ≈1 eV. Finally, making both of these changes at the same time (*i.e.*, $C_4H_2O_2S \rightarrow C_4H_3N$) leads to further changes in both $\langle\alpha\rangle$ and $\langle E_{gap}\rangle$. While the last segment on this front can be rationalized *a posteriori* as a series of structural/compositional changes resulting in more compact and less conjugated molecules that exhibit reduced $\langle\alpha\rangle$ and increased $\langle E_{gap}\rangle$ values, the *a priori* prediction of these molecules is far from trivial.

After performing a similar analysis on the QM7-X molecules with $\langle E_{AT}\rangle \in [60,70)$ eV and $\langle E_{AT}\rangle \in [70,80)$ eV, we again found numerous examples of how the intrinsic flexibility woven into CCS manifests during the molecular design process (see Fig. 4). In the $\langle E_{AT}\rangle \in [60,70)$ eV sector, the Ⓒ → Ⓓ front forms an effectively straight line bisecting ($\langle\alpha\rangle$, $\langle E_{gap}\rangle$)-space, indicating that most of its 12 constituent molecules correspond to Pareto-optimal solutions which simultaneously optimize both $\langle\alpha\rangle$ and $\langle E_{gap}\rangle$. Notable exceptions worth mention include the sixth, seventh, and eighth molecules on this front (*i.e.*, C_6H_6O, C_7H_4, C_6H_6), wherein we first observe a sharp increase in $\langle E_{gap}\rangle$ accompanied by almost no change in $\langle\alpha\rangle$ as C_6H_6O (a kinked molecule with two alkynes connected by a central alcohol group) is transformed into C_7H_4 (a propeller-like molecule with three terminal alkynes connected by a central aliphatic (CH) group). As mentioned above, such changes can often be rationalized retrospectively, *i.e.*, here we would argue that the broken conjugation (among the three triple bonds) in C_7H_4 will further localize the π-electrons, thereby increasing $\langle E_{gap}\rangle$. In the same breath, it is perhaps less straightforward to rationalize why C_6H_6O and C_7H_4 have nearly identical $\langle\alpha\rangle$ values, besides making the somewhat hand-waving argument that these molecules have similar molecular volumes. This was followed by a sharp decrease in $\langle\alpha\rangle$ accompanied by almost no change in $\langle E_{gap}\rangle$ as C_7H_4 is transformed into C_6H_6 (a staggered molecule with two terminal alkynes connected by a central ethylene (–CH_2–CH_2–) group). In this case, one could argue (with some conviction) that such a transition does not affect the overall mobility of the π-electrons (hence no appreciable change in $\langle E_{gap}\rangle$); in the same breath, this transition clearly involves the loss of a C atom and the gain of two H atoms, which will tend to decrease $\langle\alpha\rangle$ since $\tilde{\alpha}_C > 2\tilde{\alpha}_H$. In the $\langle E_{AT}\rangle \in [70,80)$ eV sector, the Ⓔ → Ⓕ front is more parabolic in shape and primarily consists of structural/constitutional isomers punctuated by simple functional group changes; for brevity, we leave a more detailed analysis of this front to the interested reader.

4 Summary and outlook

In this work, we used the recently developed QM7-X dataset—which includes 42 physicochemical properties obtained *via* high-level QM calculations for ≈4.2M (equilibrium and non-equilibrium) molecular structures containing up to seven heavy atoms—to study the structure–property/property–property relationships (SPR/PPR) that exist in the sector of chemical compound space (CCS) spanned by small (primarily organic) molecules. By quantitatively analyzing the pairwise correlations between a number of diverse physicochemical properties, we found that a vast majority (>90%) exhibited little (to no) correlation. While one might view this as a challenge for rational molecular design, we argued that this finding highlights an intrinsic flexibility—or "freedom of design"—that exists throughout CCS, as there are generally no hard-and-fast limitations preventing a molecule from exhibiting any pair of these properties.

We then investigated how this intrinsic degree of flexibility depends on the structure and chemical composition (*i.e.*, the tunable knobs underlying the molecular design process) of the molecules in this sector of CCS. Through a series of detailed case studies, we showcased numerous examples of structurally and/or compositionally distinct molecules that share multiple properties, demonstrating that the "freedom of design" conjecture proposed herein applies rather generally to both extensive and intensive properties, independent of whether such properties exhibit strong or weak mutual correlations. We also showed that these findings are quite robust and effectively independent of whether one works with normalized (or non-normalized) variants of the extensive properties. Quite interestingly, this analysis also indicated that the choice of normalization protocol might be another knob that can be used to selectively enhance/optimize the flexibility in CCS when searching for molecules with targeted properties.

Since molecular design often involves the simultaneous optimization of multiple properties, we also used the extensive QM7-X structure–property dataset to characterize and enumerate progressively more complex manifolds of molecular property space. By performing a tiered multi-property analysis involving exhaustive enumeration over the ≈41k equilibrium molecules in QM7-X, we found a remarkably large number of structurally and/or compositionally distinct molecules that share multiple properties—yet another manifestation of "freedom of design" in CCS. For instance, we were able to find ≈4k unique 4-molecule sets that share the same E_{AT}, α, E_{gap}, and μ values as well as ≈10k unique 4-molecule sets that share the same E_{AT}, α, E_{HOMO}, and E_{LUMO} values. While a number of different extensive, intensive, ground-state, and response properties were included in our analysis, additional research will be needed to assess the full extent of "freedom of design" when considering more advanced (*i.e.*, optical, excited-state) properties across larger swaths of CCS.

To explore how this intrinsic flexibility will manifest in the molecular design process, we then used Pareto multi-property optimization to search for molecules in QM7-X with simultaneously large α and E_{gap} values. Analysis of the resulting Pareto fronts identified unique and non-trivial paths through CCS consisting of molecules connected by structural and/or compositional changes that yield simultaneously optimal values for these properties. While consecutive molecules on a given front can often be rationalized *a posteriori* using chemical/physical intuition, several interesting and unexpected molecules appeared in this analysis, reflecting the freedom one

has in the rational design and discovery of molecules with targeted property values. 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 those in QM7-X) that are missing in each front.[82,83]

By demonstrating that "freedom of design" is a fundamental and emergent property of CCS, this work has a number of important implications in the field of rational molecular design. For one, we hope this work will challenge the greater chemical sciences community to consider how such intrinsic flexibility can be used to extend the dominant paradigm in the *forward* molecular design process (*i.e.*, functional group modification based on a largely fixed molecular scaffold(s)). We also hope that this work will emphasize the critical importance of high-quality QM structure–property data in the training and development of next-generation ML approaches that will enable the exploration and characterization of more vast swaths of CCS. Additionally, this work estimated that 7–10 properties are needed to *uniquely* identify the small (primarily organic) molecules in QM7-X. This quantity is tantamount to the effective dimensionality associated with each molecule (in molecular property space) and is needed to define the *inverse* molecular design problem, in which one seeks to find a molecule (or set of molecules) corresponding to a targeted/pre-defined array of properties. At the current time, substantive progress towards solving this "holy grail" of molecular design seems possible with the use of invertible ML architectures (*e.g.*, as accomplished by generative and diffusion models[84–86]) in conjunction with diverse molecular datasets (*e.g.*, QM7-X and beyond).

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.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

LMS, JH, and AT acknowledge financial support from the European Research Council (ERC-CoG grant BeStMo). This material is based upon work supported by the National Science Foundation under Grant No. CHE-1945676. 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.

References

1 K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev and A. Walsh, *Nature*, 2018, **559**, 547–555.
2 B. Huang and O. A. von Lilienfeld, *Chem. Rev.*, 2021, **121**, 10001–10036.
3 O. A. von Lilienfeld, K.-R. Müller and A. Tkatchenko, *Nat. Rev. Chem.*, 2020, **4**, 347–358.
4 A. Tkatchenko, *Nat. Commun.*, 2020, **11**, 4125.
5 A. P. Bartók, S. De, C. Poelking, N. Bernstein, J. R. Kermode, G. Csányi and M. Ceriotti, *Sci. Adv.*, 2017, **3**, 1–8.
6 V. L. Deringer, A. P. Bartók, N. Bernstein, D. M. Wilkins, M. Ceriotti and G. Csányi, *Chem. Rev.*, 2021, **121**, 10073–10141.
7 J. A. Keith, V. Vassilev-Galindo, B. Cheng, S. Chmiela, M. Gastegger, K.-R. Müller and A. Tkatchenko, *Chem. Rev.*, 2021, **121**, 9816–9872.
8 M. Meuwly, *Chem. Rev.*, 2021, **121**, 10218–10239.
9 E. N. Muratov, J. Bajorath, R. P. Sheridan, I. V. Tetko, D. Filimonov, V. Poroikov, T. I. Oprea, I. I. Baskin, A. Varnek, A. Roitberg, O. Isayev, S. Curtalolo, D. Fourches, Y. Cohen, A. Aspuru-Guzik, D. A. Winkler, D. Agrafiotis, A. Cherkasov and A. Tropsha, *Chem. Soc. Rev.*, 2020, **49**, 3525–3564.
10 A. Cherkasov, E. N. Muratov, D. Fourches, A. Varnek, I. I. Baskin, M. Cronin, J. Dearden, P. Gramatica, Y. C. Martin, R. Todeschini, V. Consonni, V. E. Kuz'min, R. Cramer, R. Benigni, C. Yang, J. Rathman, L. Terfloth, J. Gasteiger, A. Richard and A. Tropsha, *J. Med. Chem.*, 2014, **57**, 4977–5010.
11 G. Lambrinidis and A. Tsantili-Kakoulidou, *Expert Opin. Drug Discovery*, 2018, **13**, 851–859.
12 R. D. Clark and P. R. Daga, in *Building a Quantitative Structure-Property Relationship (QSPR) Model*, ed. R. S. Larson and T. I. Oprea, Springer New York, New York, NY, 2019, pp. 139–159.
13 K. Roy, S. Kar and R. Das, *A Primer on QSAR/QSPR Modeling: Fundamental Concepts*, Springer International Publishing, 2015.
14 T. Le, V. C. Epa, F. R. Burden and D. A. Winkler, *Chem. Rev.*, 2012, **112**, 2889–2919.
15 J. Vamathevan, D. Clark, P. Czodrowski, I. Dunham, E. Ferran, G. Lee, B. Li, A. Madabhushi, P. Shah, M. Spitzer and S. Zhao, *Nat. Rev. Drug Discovery*, 2019, **18**, 463–477.
16 V. O. Gawriljuk, D. H. Foil, A. C. Puhl, K. M. Zorn, T. R. Lane, O. Riabova, V. Makarov, A. S. Godoy, G. Oliva and S. Ekins, *J. Chem. Inf. Model.*, 2021, **61**, 3804–3813.
17 J. M. Stokes, K. Yang, K. Swanson, W. Jin, A. Cubillos-Ruiz, N. M. Donghia, C. R. MacNair, S. French, L. A. Carfrae, Z. Bloom-Ackermann, V. M. Tran, A. Chiappino-Pepe, A. H. Badran, I. W. Andrews, E. J. Chory, G. M. Church, E. D. Brown, T. S. Jaakkola, R. Barzilay and J. J. Collins, *Cell*, 2020, **180**, 688–702.
18 T. Williams, K. McCullough and J. A. Lauterbach, *Chem. Mater.*, 2020, **32**, 157–165.
19 D. Xue, P. V. Balachandran, J. Hogden, J. Theiler, D. Xue and T. Lookman, *Nat. Commun.*, 2016, **7**, 11241.

20 H. C. Herbol, W. Hu, P. Frazier, P. Clancy and M. Poloczek, *npj Comput. Mater.*, 2018, **4**, 51.
21 L. Tallorin, J. Wang, W. E. Kim, S. Sahu, N. M. Kosa, P. Yang, M. Thompson, M. K. Gilson, P. I. Frazier, M. D. Burkart and N. C. Gianneschi, *Nat. Commun.*, 2018, **9**, 5253.
22 J.-L. Reymond and M. Awale, *ACS Chem. Neurosci.*, 2012, **3**, 649–657.
23 L. Ruddigkeit, R. van Deursen, L. C. Blum and J.-L. Reymond, *J. Chem. Inf. Model.*, 2012, **52**, 2864–2875.
24 L. C. Blum and J.-L. Reymond, *J. Am. Chem. Soc.*, 2009, **131**, 8732–8733.
25 T. Fink, H. Bruggesser and J.-L. Reymond, *Angew. Chem., Int. Ed.*, 2005, **44**, 1504–1508.
26 T. Fink and J.-L. Reymond, *J. Chem. Inf. Model.*, 2007, **47**, 342–353.
27 G. Montavon, M. Rupp, V. Gobre, A. Vazquez-Mayagoitia, K. Hansen, A. Tkatchenko, K.-R. Müller and O. A. von Lilienfeld, *New J. Phys.*, 2013, **15**, 095003.
28 M. Rupp, A. Tkatchenko, K.-R. Müller and O. A. von Lilienfeld, *Phys. Rev. Lett.*, 2012, **108**, 058301.
29 R. Ramakrishnan, P. O. Dral, M. Rupp and O. A. von Lilienfeld, *Sci. Data*, 2014, **1**, 140022.
30 J. S. Smith, O. Isayev and A. E. Roitberg, *Sci. Data*, 2017, **4**, 170193.
31 J. S. Smith, O. Isayev and A. E. Roitberg, *Chem. Sci.*, 2017, **8**, 3192–3203.
32 Y. Yang, K. U. Lao, D. M. Wilkins, A. Grisafi, M. Ceriotti and R. A. DiStasio Jr, *Sci. Data*, 2019, **6**, 152.
33 J. S. Smith, R. Zubatyuk, B. Nebgen, N. Lubbers, K. Barros, A. E. Roitberg, O. Isayev and S. Tretiak, *Sci. Data*, 2020, **7**, 134.
34 J. Hoja, L. Medrano Sandonas, B. G. Ernst, A. Vazquez-Mayagoitia, R. A. DiStasio Jr and A. Tkatchenko, *Sci. Data*, 2021, **8**, 43.
35 G. Seifert, D. Porezag and T. Frauenheim, *Int. J. Quantum Chem.*, 1996, **58**, 185–192.
36 M. Gaus, Q. Cui and M. Elstner, *J. Chem. Theory Comput.*, 2011, **7**, 931–948.
37 A. Tkatchenko, R. A. DiStasio Jr, R. Car and M. Scheffler, *Phys. Rev. Lett.*, 2012, **108**, 236402.
38 M. Stöhr, G. S. Michelitsch, J. C. Tully, K. Reuter and R. J. Maurer, *J. Chem. Phys.*, 2016, **144**, 151101.
39 J. P. Perdew, M. Ernzerhof and K. Burke, *J. Chem. Phys.*, 1996, **105**, 9982–9985.
40 C. Adamo and V. Barone, *J. Chem. Phys.*, 1999, **110**, 6158–6170.
41 A. Ambrosetti, A. M. Reilly, R. A. DiStasio Jr and A. Tkatchenko, *J. Chem. Phys.*, 2014, **140**, 18A508.
42 V. Havu, V. Blum, P. Havu and M. Scheffler, *J. Comput. Phys.*, 2009, **228**, 8367–8379.
43 V. Blum, R. Gehrke, F. Hanke, P. Havu, V. Havu, X. Ren, K. Reuter and M. Scheffler, *Comput. Phys. Commun.*, 2009, **180**, 2175–2196.
44 X. Ren, P. Rinke, V. Blum, J. Wieferink, A. Tkatchenko, A. Sanfilippo, K. Reuter and M. Scheffler, *New J. Phys.*, 2012, **14**, 053020.
45 K. Deb, A. Pratap, S. Agarwal and T. Meyarivan, *IEEE Trans. Evol. Comput.*, 2002, **6**, 182–197.
46 V. Bhaskar, S. K. Gupta and A. K. Ray, *Rev. Chem. Eng.*, 2000, **16**, 1–54.
47 J. Blank and K. Deb, *IEEE Access*, 2020, **8**, 89497–89509.
48 J. Mandal, S. Mukhopadhyay and P. Dutta, *Multi-Objective Optimization: Evolutionary to Hybrid Framework*, Springer Singapore, Singapore, 2018.
49 G. Rangaiah and A. Bonilla-Petriciolet, *Multi-Objective Optimization in Chemical Engineering: Developments and Applications*, John Wiley & Sons, Ltd, United Kingdom, 2013.
50 M. D. Hager, B. Esser, X. Feng, W. Schuhmann, P. Theato and U. S. Schubert, *Adv. Mater.*, 2020, **32**, 2000587.
51 J. Lopez, D. G. Mackanic, Y. Cui and Z. Bao, *Nat. Rev. Mater.*, 2019, **4**, 312–330.
52 H. B. G. Casimir and D. Polder, *Phys. Rev.*, 1948, **73**, 360–372.
53 K. T. Tang and M. Karplus, *Phys. Rev.*, 1968, **171**, 70–74.
54 A. Tkatchenko and M. Scheffler, *Phys. Rev. Lett.*, 2009, **102**, 073005.
55 R. A. DiStasio Jr, V. V. Gobre and A. Tkatchenko, *J. Phys.: Condens. Matter*, 2014, **26**, 213202.
56 A. Tkatchenko, A. Ambrosetti and R. A. DiStasio Jr, *J. Chem. Phys.*, 2013, **138**, 074106.
57 A. Stone, *The Theory of Intermolecular Forces*, Clarendon Press, United Kingdom, 1996.
58 R. A. DiStasio Jr, O. A. von Lilienfeld and A. Tkatchenko, *Proc. Natl. Acad. Sci. U. S. A.*, 2012, **109**, 14791–14795.
59 Y. S. Al-Hamdani and A. Tkatchenko, *J. Chem. Phys.*, 2019, **150**, 010901.
60 T. Brinck, J. S. Murray and P. Politzer, *J. Chem. Phys.*, 1993, **98**, 4305–4306.
61 S. A. Blair and A. J. Thakkar, *J. Chem. Phys.*, 2014, **141**, 074306.
62 Y.-Q. Zhao, Y. Cheng, C.-E. Hu, B.-R. Yu and G.-F. Ji, *Theor. Chem. Acc.*, 2021, **140**, 51.
63 X.-B. Li, H.-Y. Wang, R. Lv, W.-D. Wu, J.-S. Luo and Y.-J. Tang, *J. Phys. Chem. A*, 2009, **113**, 10335–10342.
64 I. Vasiliev, S. Öğüt and J. R. Chelikowsky, *Phys. Rev. Lett.*, 1997, **78**, 4805–4808.
65 A. P. Jones, J. Crain, V. P. Sokhan, T. W. Whitfield and G. J. Martyna, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 2013, **87**, 144103.
66 P. Szabó, S. Góger, J. Charry, M. R. Karimpour, D. V. Fedorov and A. Tkatchenko, *Phys. Rev. Lett.*, 2022, **128**, 070602.
67 R. G. Pearson, *J. Chem. Educ.*, 1968, **45**, 581.
68 R. G. Pearson, *J. Chem. Educ.*, 1968, **45**, 643.
69 R. G. Parr and R. G. Pearson, *J. Am. Chem. Soc.*, 1983, **105**, 7512–7516.
70 P. Politzer, P. Jin and J. S. Murray, *J. Chem. Phys.*, 2002, **117**, 8197–8202.
71 F. Meyers, S. R. Marder, B. M. Pierce and J. L. Bredas, *J. Am. Chem. Soc.*, 1994, **116**, 10703–10714.
72 P. K. Chattaraj, P. Fuentealba, P. Jaque and A. Toro-Labbé, *J. Phys. Chem. A*, 1999, **103**, 9307–9312.
73 D. M. Wilkins, A. Grisafi, Y. Yang, K. U. Lao, R. A. DiStasio Jr and M. Ceriotti, *Proc. Natl. Acad. Sci. U. S. A.*, 2019, **116**, 3401–3406.

74 M. Brieger, A. Renn, A. Sodeik and A. Hese, *Chem. Phys.*, 1983, **75**, 1–9.
75 A. D. Buckingham, in *Permanent and Induced Molecular Moments and Long-Range Intermolecular Forces*, John Wiley & Sons, Ltd, United Kingdom, 1967, ch. 2, vol. 12, pp. 107–142.
76 D. V. Fedorov, M. Sadhukhan, M. Stöhr and A. Tkatchenko, *Phys. Rev. Lett.*, 2018, **121**, 183401.
77 K. U. Lao, Y. Yang and R. A. DiStasio Jr, *Phys. Chem. Chem. Phys.*, 2021, **23**, 5773–5779.
78 Y. Yang, K. U. Lao and R. A. DiStasio Jr, *Phys. Rev. Lett.*, 2019, **122**, 026001.
79 A. H. Farmahini, S. Krishnamurthy, D. Friedrich, S. Brandani and L. Sarkisov, *Chem. Rev.*, 2021, **121**, 10666–10741.
80 K. M. Jablonka, G. M. Jothiappan, S. Wang, B. Smit and B. Yoo, *Nat. Commun.*, 2021, **12**, 2312.
81 T. Erps, M. Foshey, M. K. Luković, W. Shou, H. H. Goetzke, H. Dietsch, K. Stoll, B. von Vacano and W. Matusik, *Sci. Adv.*, 2021, **7**, eabf7435.
82 J. P. Janet, S. Ramesh, C. Duan and H. J. Kulik, *ACS Cent. Sci.*, 2020, **6**, 513–524.
83 Z. del Rosario, M. Rupp, Y. Kim, E. Antono and J. Ling, *J. Chem. Phys.*, 2020, **153**, 024112.
84 R. Gómez-Bombarelli, J. N. Wei, D. Duvenaud, J. M. Hernández-Lobato, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T. D. Hirzel, R. P. Adams and A. Aspuru-Guzik, *ACS Cent. Sci.*, 2018, **4**, 268–276.
85 E. Hoogeboom, V. G. Satorras, C. Vignac and M. Welling, Equivariant Diffusion for Molecule Generation in 3D, *arXiv*, 2022, preprint, arXiv:2203.17003, DOI: **10.48550/arXiv.2203.17003**.
86 D. M. Anstine and O. Isayev, *J. Am. Chem. Soc.*, 2023, **145**, 8736–8750.