Electronic Supplementary Information (ESI) for:

"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,[a,b] Brian G. Ernst,[c] Alvaro Vazquez-Mayagoitia,[d] Robert A. DiStasio Jr.,*[c] and Alexandre Tkatchenko*[a]

[a] *Physics and Materials Science Research Unit, University of Luxembourg, L-1511 Luxembourg, Luxembourg. Tel: +352 46 66 44 5138; E-mail: 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*
[d] *Computational Science Division, Argonne National Laboratory, Argonne, IL 60439, USA*

* Corresponding authors: Leonardo Medrano Sandonas (`leonardo.medrano@uni.lu`), Robert A. DiStasio Jr. (`distasio@cornell.edu`), Alexandre Tkatchenko (`alexandre.tkatchenko@uni.lu`)

Contents

1 Influence of non-equilibrium molecular structures on pairwise correlations in molecular property space

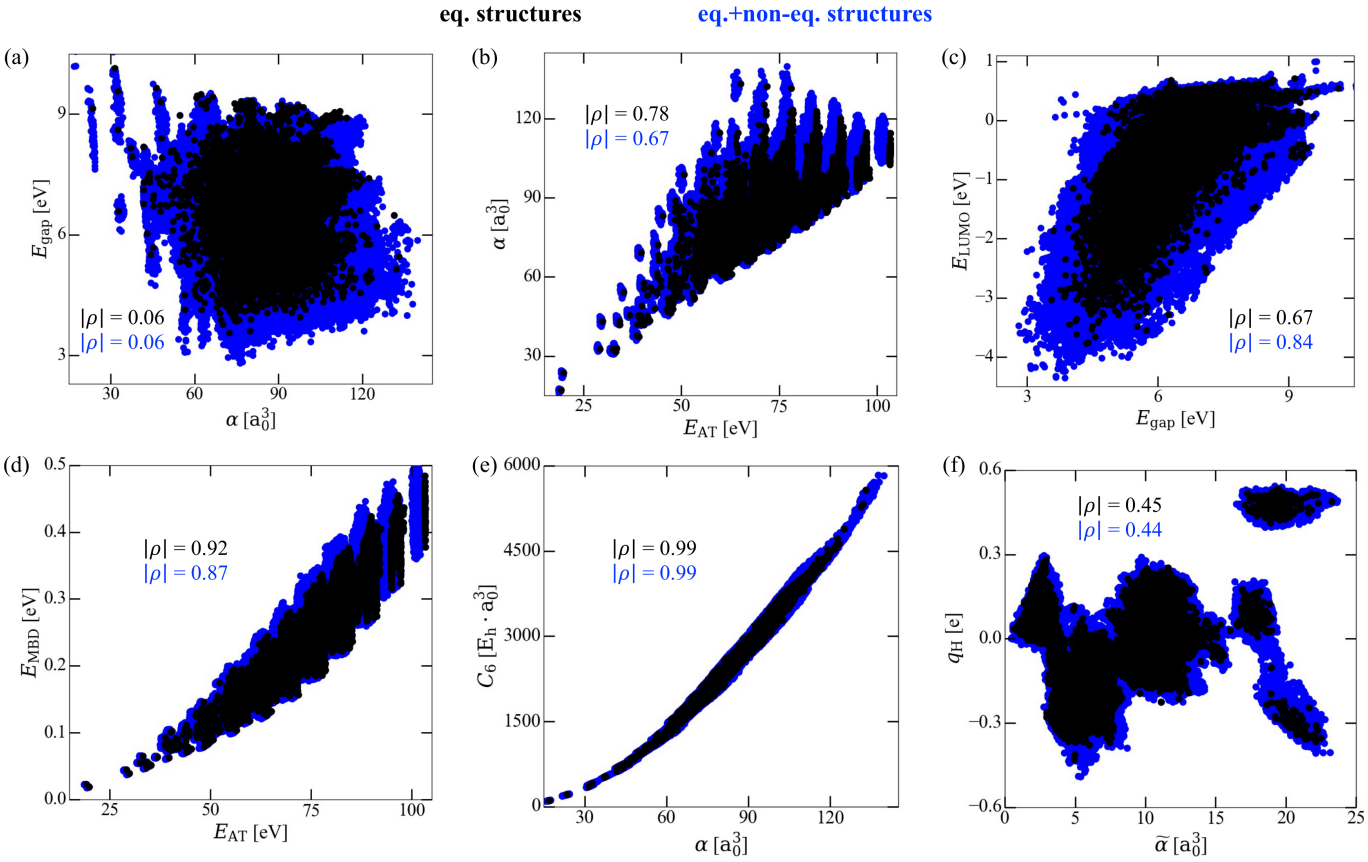

Fig. S1 Select 2D projections of the 42D QM7-X molecular property space for: (a) E_{gap} and α, (b) α and E_{AT}, (c) E_{LUMO} and E_{gap}, (d) E_{MBD} and E_{AT}, (e) C_6 and α, and (f) q_H and $\tilde{\alpha}$ (see Table 1 in the main text for a detailed description of each property). The black points represent the ≈ 41 k equilibrium structures in QM7-X, while the blue points represent all ≈ 4.2 M (equilibrium and non-equilibrium) structures in this dataset. Pearson correlation coefficients $|\rho|$ (see Eq. (1) in the main text) are also provided for each 2D projection (black font: equilibrium structures; blue font: equilibrium and non-equilibrium structures). While the inclusion of non-equilibrium molecular structures can increase or decrease the computed $|\rho|$ values, such changes do not substantially alter the degree of correlation between properties.

Extended Discussion. As alluded to in Sec. 3.1 of the main text, there were a handful of cases in the moderately and strongly correlated sectors that underwent more substantive changes when $|\rho|$ was computed using the equilibrium structures only. For example, both (E_{AT}, C_6) and (E_{AT}, α) have $|\rho|$ values that increase by ≈ 0.10 (but remain below 0.91) when computed using the equilibrium structures only (see Fig. S1(b)); as such, these PPR would still be classified as moderately correlated according to the original scheme. Among the moderately and strongly correlated PPR, only one case crossed the boundary between these classification categories, *i.e.*, (E_{AT}, E_{MBD}), for which $|\rho|$ increased by ≈ 0.05 (from 0.87 to 0.92) when computed using the equilibrium structures only (see Fig. S1(d)); since $|\rho|$ started at the higher end of the moderately correlated sector and the change in $|\rho|$ was not substantial, we take this as further evidence that our working classification system is fairly robust.

2 Influence of non-equilibrium molecular structures on the distribution of extensive and intensive property values

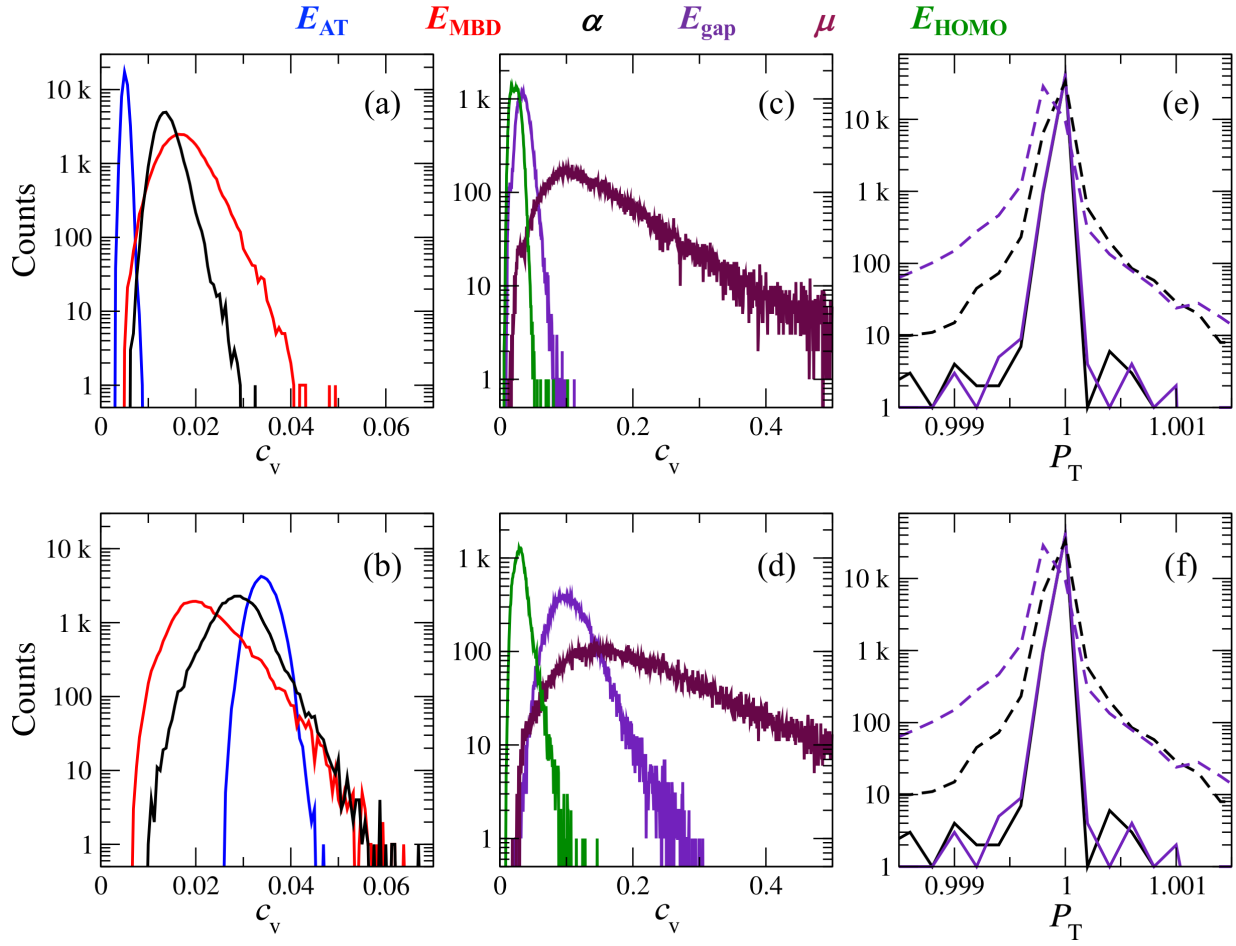

Fig. S2 Influence of non-equilibrium conformations on the coefficient of variation ($c_v = \sigma_x/\bar{x}$, the ratio of the standard deviation σ_x to the mean value $\bar{x}$ for a given property x) in a representative set of: (a)–(b) extensive properties (E_{AT}, E_{MBD}, α) and (c)–(d) intensive properties (E_{gap}, μ, E_{HOMO}). Also depicted is the influence of the non-equilibrium conformations on thermally (Boltzmann) averaged property values as measured by $P_T \equiv \langle x \rangle_T / x_{eq}$ (the ratio of the thermally-averaged value $\langle x \rangle_T$ to the value computed using the equilibrium structure only x_{eq}) for: (e)–(f) α and E_{gap} at $T = 300$ K (solid lines) and $T = 900$ K (dashed lines). The effects of including the 50 lowest-energy non-equilibrium conformations (per equilibrium structure) when computing c_v and P_T are depicted in panels (a), (c), and (e); the effects of including all 100 non-equilibrium conformations are depicted in panels (b), (d), and (f). From panels (a)–(d), one can see that the inclusion of non-equilibrium conformations tends to have a larger effect on the c_v values corresponding to the intensive properties (rather than the extensive properties). The plots in panels (e) and (f) show that Boltzmann averaging at $T = 300$ K has very little influence on the distribution of both intensive and extensive properties (*i.e.*, $P_{300\text{ K}} \approx 1$); more pronounced differences were observed at $T = 900$ K, as higher-energy (*i.e.*, more distorted) structures are more heavily weighted at such elevated temperatures.

3 Influence of molecular size and chemical composition on the distribution of extensive and intensive property values

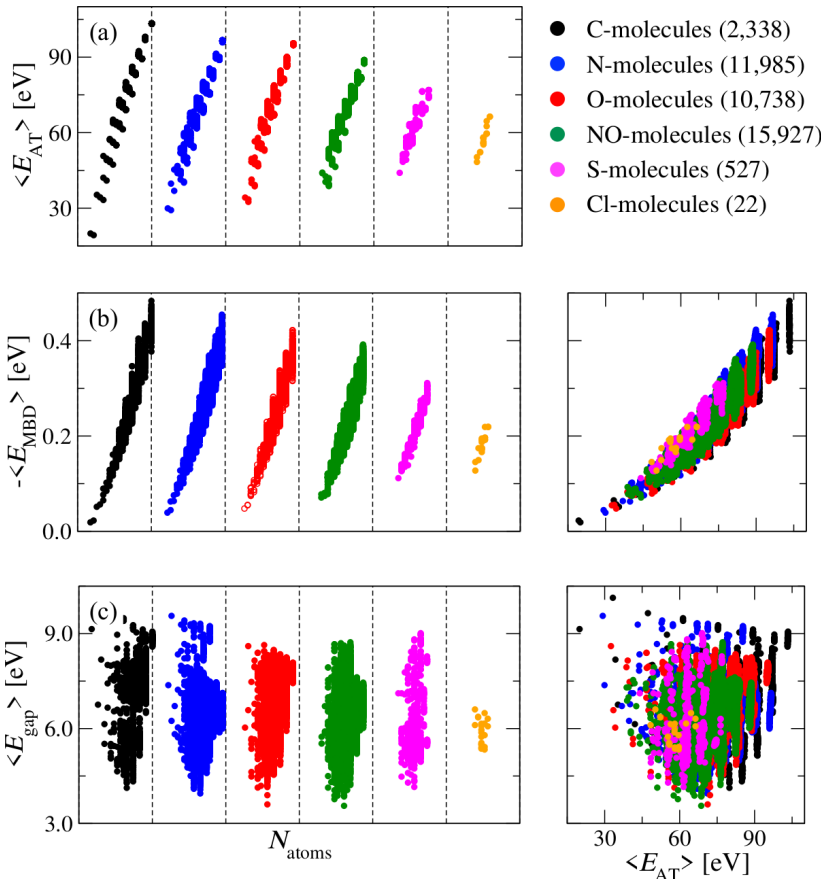

Fig. S 3 Influence of molecular size (as measured by the total number of atoms N_{atoms}) and chemical composition in a representative set of thermally-averaged: (a)–(b) extensive properties ($\langle E_{\text{AT}} \rangle$ and $-\langle E_{\text{MBD}} \rangle$) and (c) intensive properties ($\langle E_{\text{gap}} \rangle$) obtained by Boltzmann averaging over all 101 (equilibrium and non-equilibrium) conformations per equilibrium structure in QM7-X at $T = 300$ K. For clarity, the x-axes in each of these panels have been split into six windows (delineated by vertical dashed lines), each of which ranges from $N_{\text{atoms}} = 0$ to $N_{\text{atoms}} = 23$ (*i.e.*, the largest molecule in QM7-X). Also provided are correlation plots depicting the variation in $-\langle E_{\text{MBD}} \rangle$ and $\langle E_{\text{gap}} \rangle$ as a function of $\langle E_{\text{AT}} \rangle$ (panels (b) and (c) only). In these plots, each point was colored according to the heteroatoms contained in each molecule (see legend in panel (a)), *i.e.*, the black points correspond to the 2,338 molecules in QM7-X that contain C atoms and no heteroatoms, while the blue points correspond to the 11,985 molecules in QM7-X that contain N atoms (and no other heteroatoms). From these plots, one can observe that: (*i*) chemical composition does not seem to have a significant effect on property-property relationships (as evidenced by the similar correlation plots in each of the six windows above) and (*ii*) when searching for distinct molecules with a desired set of properties, one would have the most flexibility when considering the medium-sized molecules in QM7-X (as evidenced by the fact that the correlation plots have the largest spread for intermediate N_{atoms} values). Here, we note that the latter finding can be significantly modified by normalizing the extensive properties (see Sec. 3.2 of the main text and Fig. S4).

4 Influence of normalization on pairwise correlations in molecular property space

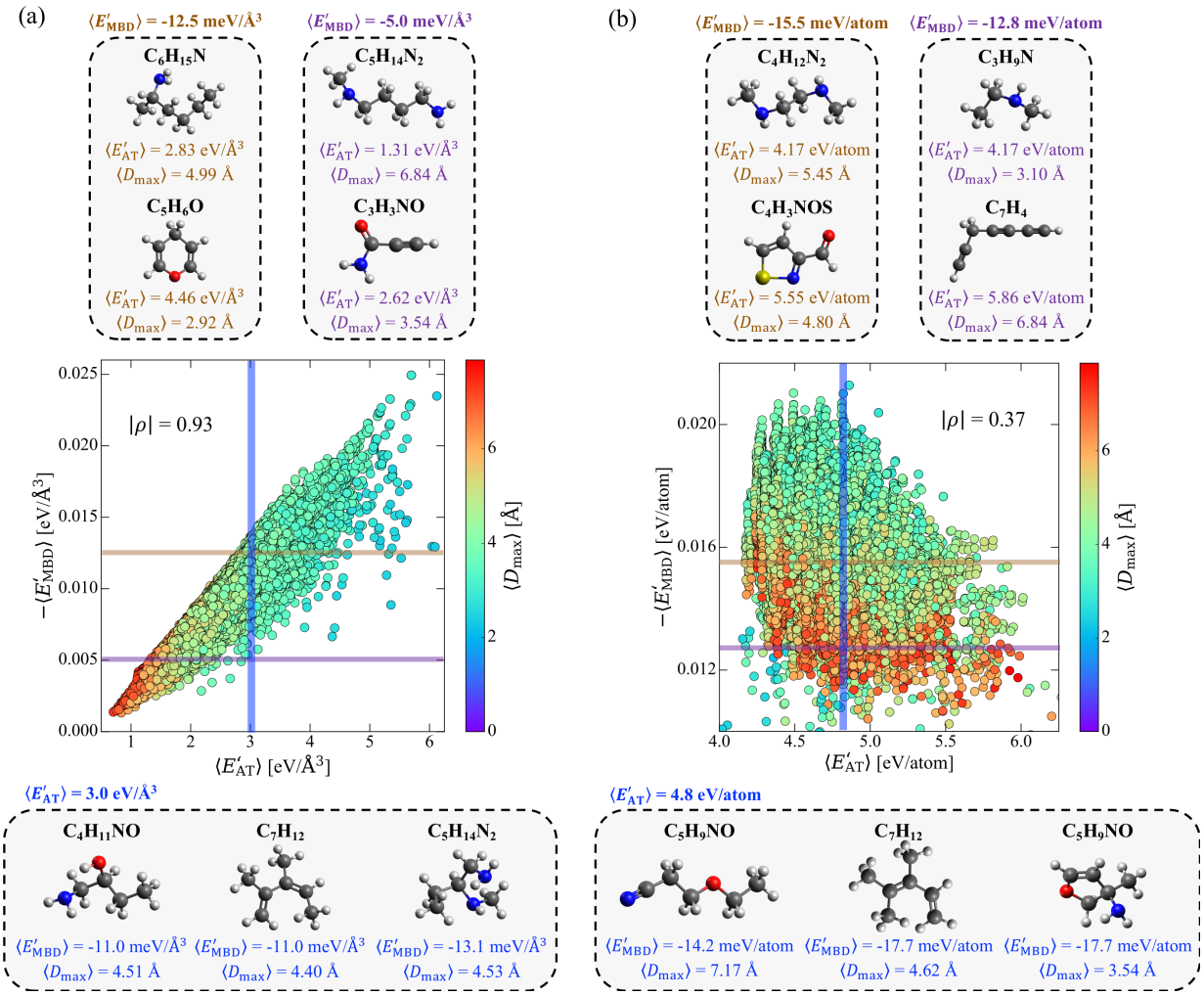

Fig. S4 Influence of normalization on the pairwise correlation between thermally-averaged (T = 300 K) extensive properties ($-\langle E_{\text{MBD}}\rangle$ and $\langle E_{\text{AT}}\rangle$). To allow for a direct comparison to Fig. 2(a) in the main text, correlation plots are provided for normalized variants of these extensive properties (*i.e.*, $\langle E_{\text{MBD}}\rangle \rightarrow \langle E'_{\text{MBD}}\rangle$ and $\langle E_{\text{AT}}\rangle \rightarrow \langle E'_{\text{AT}}\rangle$) with respect to: (a) the thermally-averaged molecular volume V and (b) the total number of atoms N_{atoms} in a given molecule. Each point in these plots has also been colored according to the corresponding thermally-averaged maximum distance between heavy/non-hydrogen atoms ($\langle D_{\text{max}}\rangle$); see Sec. 2 of the main text for more details. From the correlation plots in panels (a) and (b), one can see that the degree of correlation between $\langle E'_{\text{MBD}}\rangle$ and $\langle E'_{\text{AT}}\rangle$ strongly depends on the chosen normalization quantity and can therefore be quite variable. When normalizing these properties with respect to $\langle V\rangle$, $|\rho|$ slightly increased from 0.92 (extensive/non-normalized properties, see Fig. 2(a)) to 0.93 (panel (a)); when normalizing with respect to N_{atoms}, $|\rho|$ substantially decreased to 0.37 (panel (b)). Despite such non-trivial changes to $|\rho|$, we were still able to find—using either of these normalization protocols—structurally and/or compositionally distinct molecules with the same $\langle E'_{\text{MBD}}\rangle$ but different $\langle E'_{\text{AT}}\rangle$ (and vice versa), as well as markedly distinct molecules sharing both of these properties; these molecules are analogous to those obtained using the extensive/non-normalized variants (*cf.* the top and bottom insets in panels (a) and (b) with those in Fig. 2(a)). While these findings provide strong evidence that our "freedom of design" conjecture is quite robust, the option to work with normalized properties with a significantly reduced degree of correlation (*i.e.*, the N_{atoms}-normalized variant in panel (b)) can provide additional flexibility when searching for distinct molecules with targeted pairs of property values.

Extended Discussion. As one might expect, the dispersion in the $(\langle E'_{\mathrm{MBD}}\rangle, \langle E'_{\mathrm{AT}}\rangle)$ plot in Fig. S4(a) is no longer correlated with $\langle D_{\mathrm{max}}\rangle$, as normalization with respect to $\langle V\rangle$ explicitly accounts for the spatial extent of each molecule (thereby making $\langle D_{\mathrm{max}}\rangle$ redundant). In this case, the largest molecules in QM7-X are now relegated to the region of small $|\langle E'_{\mathrm{MBD}}\rangle|$ and $|\langle E'_{\mathrm{AT}}\rangle|$ (due to their correspondingly large $\langle V\rangle$), while the small- and medium-sized molecules are spread throughout the correlation plot. On the other hand, normalization with respect to N_{atoms} represents a more generic scaling process that does not explicitly account for the spatial extent of each molecule (*i.e.*, beyond the fact that molecules with more atoms will tend to have larger spatial extents than molecules with less atoms). In this case, normalization results in the weakly correlated $(\langle E'_{\mathrm{MBD}}\rangle, \langle E'_{\mathrm{AT}}\rangle)$ plot depicted in Fig. S4(b), which effectively resembles a structure-less "blob" (albeit) with dispersion in $\langle E'_{\mathrm{MBD}}\rangle$ that is still fairly well-correlated with $\langle D_{\mathrm{max}}\rangle$. Working with such normalized properties (*i.e.*, that have a significantly reduced degree of correlation) can provide additional flexibility when searching for structurally and/or compositionally distinct molecules with targeted property values; how the choice of normalization protocol can be used to enhance (or perhaps optimize) the "freedom of design" in CCS is clearly a question of fundamental importance that warrants more thorough examination and is beyond the scope of this work.

5 Multi-property analysis in molecular property space: Influence of HOMO and LUMO energies

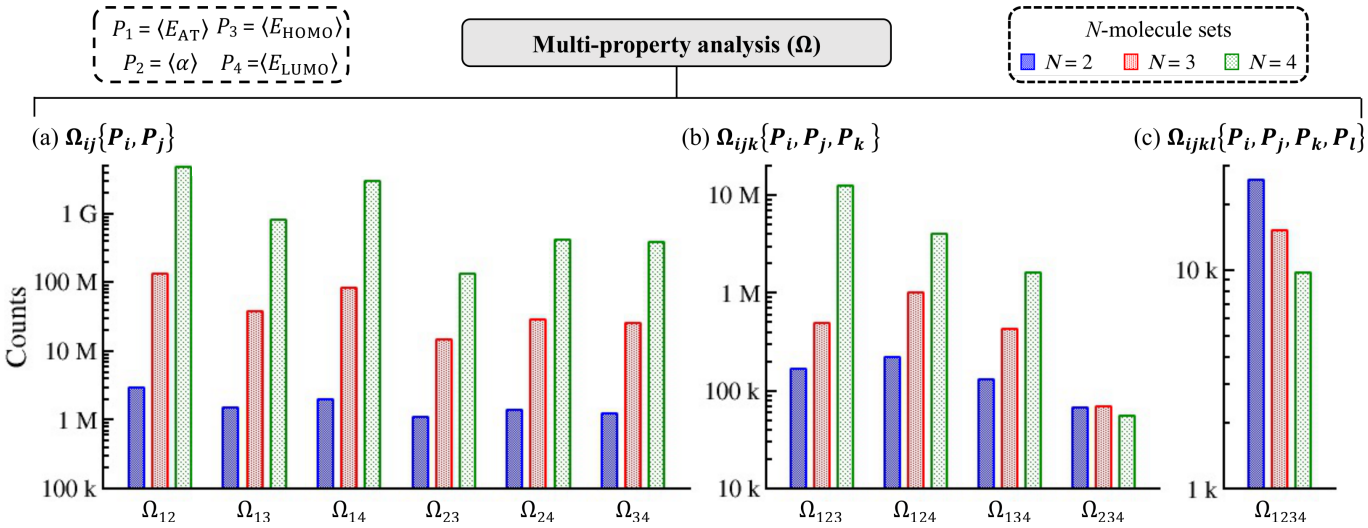

Fig. S5 Three different multi-property analyses (Ω) were performed to exhaustively enumerate the number of N-molecule sets (*i.e.*, sets containing $N = 2, 3, 4$ unique molecules taken from the $\approx$ 41 k equilibrium molecules in QM7-X) that share: **(a)** two properties ($\Omega_{ij}\{P_i, P_j\}$), **(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 ($P_3 = \langle E_{HOMO} \rangle$ and $P_4 = \langle E_{LUMO} \rangle$) properties evaluated at $T = 300$ K. 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 larger than those shown in Fig. 3(b,c) in the main text.