In fact, the chemical subspaces that have been analyzed the most are organic ones due to the more graph theoretically tractable nature of carbon scaffolds. The usual number of molecules in the chemical space of organic molecules with less than 500 Da is estimated to be 10^{60}.?? Consequently, including molecules and materials of all sizes and shapes to cover and enumerate the whole chemical space is utterly intractable.

The largest databases of produced molecules are not only just a fraction of the actual space but is also heavily biased towards easily accessible molecules through synthesis and other biases.?? Enumeration projects of theorized molecules have attempted to either exhaust or systematically cover?? large subspaces. The GDB-17 dataset?? tries to enumerate all possible organic scaffold based structures of up to 17 atoms of C, N, O, S and halogens. This results in approximately 166 billion organic small molecules, exhibiting more diversity than other data sets and lead to new discoveries.?? Instead of going through all possibilities of structures, Virshup *et al.*?? introduced an algorithm to stochastically?? sample the chemical space for what they call "representative sublibraries, maximally diverse representative collections of compounds that contain as much diversity as the parent library expressed in a much smaller number of compounds." With this method, a particular subspace can be accurately represented by much less molecules.

Experimentally, the emergence of (virtual) high-throughput chemistry has lead to successes and might give some remedy to the intractability of the full space.(REF)

-Computational high-throughput screening is key in chemical and materials discovery,111 but high computational cost has limited chemical space exploration to a small fraction of feasible compounds.12,13 1 HT (curtarolo): https://www.nature.com/articles/nmat3568 2 HT (Norskov): https://www.nature.com/articles/nmat1752 3 Cat design (Norskov): https://www.nature.com/articles/nchem.121 4 Acc DFT w GA (Vegge): https://pubs.acs.org/doi/10.1021/acs.chemmater.5b00446 5 ML x DFT (Ceder): https://pubs.acs.org/doi/10.1021/cm100795d 6 Databases (Ceder): https://www.sciencedirect.com/science/article/pii/S0927025611001133?via7 oxides (Hautier/Ceder): https://www.nature.com/articles/ncomms3292 8 exp: https://pubs.acs.org/doi/10.1021/acs.inorgchem.5b01409 9 fingerprints (bajorath): https://www.sciencedirect.com/science/article/pii/S1359644607000529?via10 clean ene (guzik): https://pubs.acs.org/doi/10.1021/jz200866s 11 diodes: https://aip.scitation.org/doi/10.1063 found: https://www.annualreviews.org/doi/10.1146/annurev.matsci.38.060407.130217 12 ¿¿what i want (beratan) https://pubs.acs.org/doi/10.1021/ja401184g 13 prose smu: https://www.nature.com/articles/432823a

For inorganic coordination chemistry, there are currently no data bases. In this group's research program, we try to systematically explore transition metal complex (TMC) space. Transition metal complexes form promising functional inorganic materials due to their wide range of tunable electronic properties. However, exhaustive enumeration and calculation of all possible ligand fields is clearly intractable due to the vast nature of chemical space. For this unique challenge, the open-source software molSimplify?? was introduced for the rapid structure generation and discovery.

Even though TMCs are crucial for contemporary challenges, such as spin-crossover complexes,?? dye-sensitizers in solar cells,?? or open-shell catalysts?? few benchmark data sets, experimental data bases or softwares are available.

For enumeration, it is important to find the right representation of molecules. Computationally generated data sets consist of the molecule's identity and a number of descriptors. Chemical space can be defined as a Cartesian space in the dimension of the number of the features. Therefore, each set of descriptors spans chemical space in a different way including some molecules that possibly overlap if the descriptor set is not diverse enough. Our descriptors are introduced in section X.

-For modest sized data sets, descriptor set selection is especially critical4244 for successful ML modeling. Good feature sets should43 be cheap to compute, be as low dimensional as possible, 42 https://journals.aps.org/prb/pdf/10.1103/PhysRevB.87.184115 43 https://journals.aps.org/prl/abstract/10.1103/PhysRevLett.114.105503 44 https://aip.scitation.org/doi/10.

-Descriptors that work well for organic molecules have proven unsuitable for inorganic materials51 or molecules.52 This lack of transferability can be readily rationalized: it is well-known5255 that some electronic properties of transition metal complexes (e.g., spin state splitting) are disproportionately sensitive to the direct ligand atom identity that dominates ligand field strength.56,57 Unlike organic molecules, few force fields have been established that can capture the full range of inorganic chemical bonding.58 With the unique challenges of inorganic chemistry65 51 https://journals.aps.org/prb/abstract/10.1103/PhysRevB.89.205118 52 ANN JP: http://pubs.rsc.org/en/Content/ArticleLanding/2017/SC/C7SC01247K ::

own a elapsite: https://journals.aps.org/prl/pdf/10.1103/PhysRevLett.117.135502 b comb x ml: https://journals.aps.org/prb/pdf/10.1103/PhysRevB.89.094104