1 Size of transition metal complex (TMC) space

Among the most extensively analyzed chemical subspaces are the organic ones. Due to the graph theoretically tractable nature of carbon scaffolds enumeration is much easier than in other subspaces. The usual number of molecules in the chemical space of organic molecules with less than 500 Da is estimated to be 10^{60}.[1, 2, 3, 4, 5] Including larger molecules and materials from the whole periodic table ends up in a vastly larger number of molecules still.

For enumeration, it is important to find the right representation of molecules.[6, 7] 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 called RACs are introduced in section X.

The largest databases of existing molecules are not only just a fraction of the actual space but is also heavily biased towards easily accessible molecules through synthesis.[8, 9, 10] Computational high-throughput screening[11, 12, 13, 14, 15, 16, 17, 18] is a potential remedy but severly constraint by computational cost.[19] Enumeration projects of computationally generated molecules have attempted to either exhaust or systematically cover[**?**] large subspaces. Most prominently, the GDB-17 dataset[20] 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 a so-called representative sublibrary, which is defined as representative subspace with a smaller amount of molecules than the parent space without losing it's diversity.

TMCs form promising functional inorganic materials due to their wide range of tunable electronic properties. They are crucial for contemporary challenges, such as spincrossover complexes,[21, 22, 23] dye-sensitizers in solar cells,[24] or open-shell catalysts.[25] Nonetheless, few benchmark data sets, experimental data bases or softwares are available. In this group's research program, we try to systematically explore the TMC space. However, exhaustive enumeration and calculation of all possible ligand fields is as intractable as in organic chemistry. For this, the open-source software molSimplify[26] was developed for the rapid structure generation in coordination chemistry.

2 The role of symmetry

As in organic data sets, the size of the set scales with the number of atoms allowed, the number of elements included and other factors. In coordination chemistry, there are four basic components that increase the size of the TMC subspace under consideration: i) geometry, ii) metal center, iii) ligand, and iv)

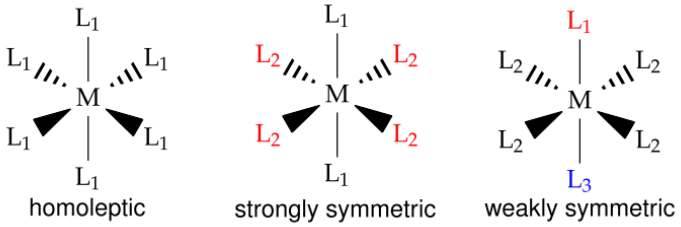

Figure 1: Examples for symmetry classes: A homoleptic complex with six times the same ligand, a strongly symmetric complex with the same ligand in axial and equatorial position, respectively, and a weakly symmetric ligand, which in addition also breaks the axial symmetry from the strongly symmetric complex.

symmetry class. Most common geometries for TMCs are octahedral, tetrahedral, or trigonal bipyramidal. The metal centers can be chosen from the transition metals from the periodic table, whereas the ligand is usually an organic molecule. These three variables scale the space by a constant linear factor, e.g. the subspace composed of the metal centers chromium and manganese is twice the size of one only composed of chromium. Component iv), the symmetry class, is usually the bottleneck variable when looking at subspaces of reasonable size.

To exemplify the scaling, we look at three different symmetry classes as shown in Figure 1. Without loss of generality we chose the octahedral geometry for simplicity. The homoleptic case will result in

$$m \cdot l \tag{1}$$

different complexes, where m and l are the number of available metal centers and ligands, respectively. For the strongly symmetric case, there will be

$$m \cdot \frac{l!}{(l-2)!} \tag{2}$$

distinguishable complexes, whereas in the weakly symmetric case we end up with

$$m \cdot \frac{l!}{(l-3)! \cdot 2} \tag{3}$$

complexes. Using four different metal centers and only 10 different ligands, we end up at 40, 360, and 1140 complexes. This exponential scaling prevents us from exhaustive analysis of the space. For lower symmetries, even enumeration becomes intractable. A complex with arbitrary ligands generates a data set with $4 \cdot 10^6$ complexes.

3 Spectrochemical Series

[27, 28, 29]

References

[1] Blair, C. M.; Henze, H. R. THE NUMBER OF STEREOISOMERIC AND NON-STEREOISOMERIC PARAFFIN HYDROCARBONS, *54,* 1538-1545.

[2] Polishchuk, P. G.; Madzhidov, T. I.; Varnek, A. Estimation of the Size of Drug-like Chemical Space Based on GDB-17 Data, *27,* 675-679.

[3] Bohacek, R. S.; McMartin, C.; Guida, W. C. The Art and Practice of Structurebased Drug Design: A Molecular Modeling Perspective, *16,* 3-50.

[4] Wester, M. J.; Pollock, S. N.; Coutsias, E. A.; Allu, T. K.; Muresan, S.; Oprea, T. I. Scaffold Topologies. 2. Analysis of Chemical Databases, *48,* 1311-1324.

[5] Triggle, D. J. The Chemist as Astronaut: Searching for Biologically Useful Space in the Chemical Universe, *78,* 217-223.

[6] Bartk, A. P.; Kondor, R.; Csnyi, G. On Representing Chemical Environments, *87,*.

[7] Ghiringhelli, L. M.; Vybiral, J.; Levchenko, S. V.; Draxl, C.; Scheffler, M. Big Data of Materials Science: Critical Role of the Descriptor, *114,* 105503.

[8] Tan, D. S. Diversity-Oriented Synthesis: Exploring the Intersections between Chemistry and Biology, *1,* 74-84.

[9] Hajduk, P. J.; Galloway, W. R. J. D.; Spring, D. R. Drug Discovery: A Question of Library Design, *470,* 42-43.

[10] Galloway, W. R. J. D.; Isidro-Llobet, A.; Spring, D. R. Diversity-Oriented Synthesis as a Tool for the Discovery of Novel Biologically Active Small Molecules, *1,* 80.

[11] Hachmann, J.; Olivares-Amaya, R.; Atahan-Evrenk, S.; Amador-Bedolla, C.; Snchez-Carrera, R. S.; Gold-Parker, A.; Vogt, L.; Brockway, A. M.; Aspuru-Guzik, A. The Harvard Clean Energy Project: Large-Scale Computational Screening and Design of Organic Photovoltaics on the World Community Grid, *2,* 2241-2251.

[12] Jain, A.; Hautier, G.; Moore, C. J.; Ping Ong, S.; Fischer, C. C.; Mueller, T.; Persson, K. A.; Ceder, G. A High-Throughput Infrastructure for Density Functional Theory Calculations, *50,* 2295-2310.

[13] Hautier, G.; Fischer, C. C.; Jain, A.; Mueller, T.; Ceder, G. Finding Natures Missing Ternary Oxide Compounds Using Machine Learning and Density Functional Theory, *22,* 3762-3767.

[14] Jensen, P. B.; Bialy, A.; Blanchard, D.; Lysgaard, S.; Reumert, A. K.; Quaade, U. J.; Vegge, T. Accelerated DFT-Based Design of Materials for Ammonia Storage, *27,* 4552-4561.

[15] Nrskov, J. K.; Bligaard, T.; Rossmeisl, J.; Christensen, C. H. Towards the Computational Design of Solid Catalysts, *1,* 37-46.

[16] Greeley, J.; Jaramillo, T. F.; Bonde, J.; Chorkendorff, I.; Nrskov, J. K. Computational High-Throughput Screening of Electrocatalytic Materials for Hydrogen Evolution, *5,* 909-913.

[17] Curtarolo, S.; Hart, G. L. W.; Nardelli, M. B.; Mingo, N.; Sanvito, S.; Levy, O. The High-Throughput Highway to Computational Materials Design, *12,* 191-201.

[18] Rajan, K. Combinatorial Materials Sciences: Experimental Strategies for Accelerated Knowledge Discovery, *38,* 299-322.

[19] Kirkpatrick, P.; Ellis, C. "Chemical Space", .

[20] Ruddigkeit, L.; van Deursen, R.; Blum, L. C.; Reymond, J.-L. Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17, *52,* 2864-2875.

[21] Ltard, J.-F.; Guionneau, P.; Goux-Capes, L. Towards Spin Crossover Applications. In *Spin Crossover in Transition Metal Compounds III*; Topics in Current Chemistry Springer, Berlin, Heidelberg:.

[22] Halcrow, M. A. Structure:Function Relationships in Molecular Spin-Crossover Complexes, *40,* 4119-4142.

[23] Ashley, D. C.; Jakubikova, E. Ironing out the Photochemical and Spin-Crossover Behavior of Fe(II) Coordination Compounds with Computational Chemistry, *337,* 97-111.

[24] Bignozzi, C. A.; Argazzi, R.; Boaretto, R.; Busatto, E.; Carli, S.; Ronconi, F.; Caramori, S. The Role of Transition Metal Complexes in Dye Sensitized Solar Devices, *257,* 1472-1492.

[25] Harvey, J. N.; Poli, R.; Smith, K. M. Understanding the Reactivity of Transition Metal Complexes Involving Multiple Spin States, *238-239,* 347-361.

[26] Ioannidis, E. I.; Gani, T. Z. H.; Kulik, H. J. molSimplify: A Toolkit for Automating Discovery in Inorganic Chemistry, *37,* 2106-2117.

[27] Tsuchida, R. Absorption Spectra of Co-Ordination Compounds. I, *13,* 388-400.

[28] Ballhausen, C. J.; Weiner, M. A. Introduction to Ligand Field Theory, *110,* 97C-97C.

[29] Griffith, J. S.; Orgel, L. E. Ligand-Field Theory, *11,* 381-393.