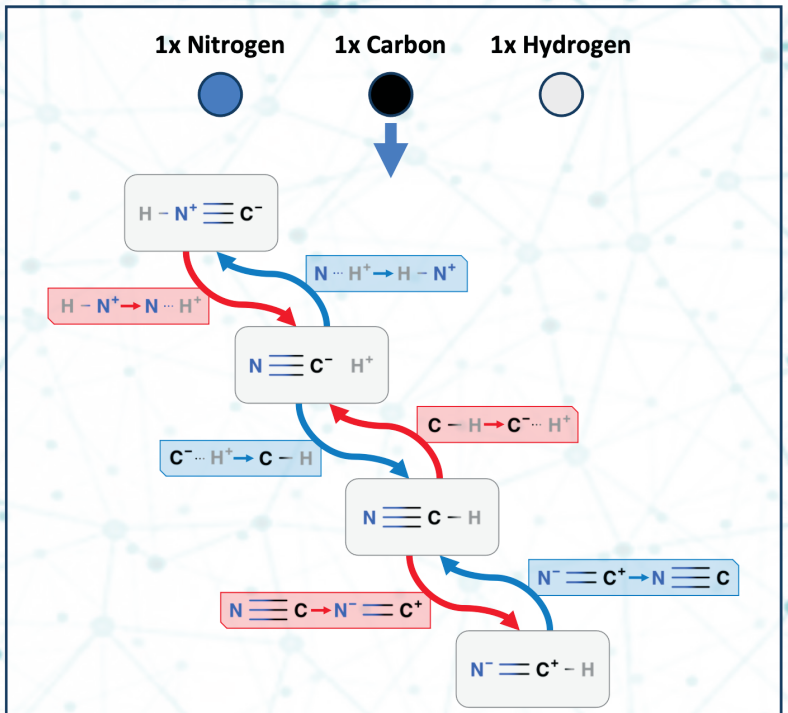
Discrete Network Representations Of Chemical Space(s)

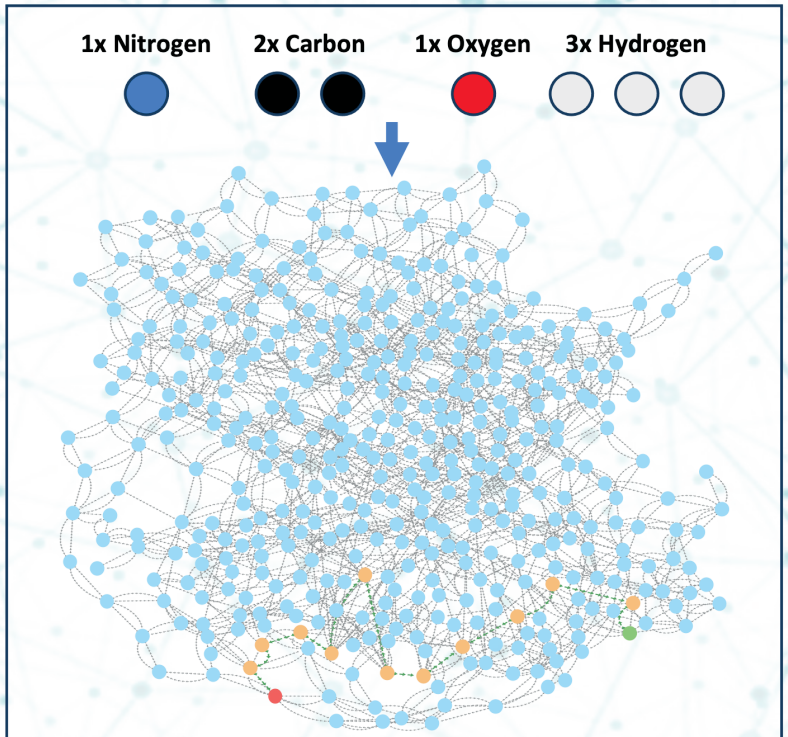
To find new molecules or predict reaction pathways between molecules, one must explore the corresponding chemical space, which is the space containing all possible chemical structures, i.e., molecules, for a set of atoms. Chemical space is extremely vast, diverse, and difficult to explore being high dimensional. The more atoms in the molecule the more difficult it is to explore the relevant chemical space as it grows exponentially with each additional atom. We develop a lower dimensional discrete network representation of chemical space in order to make it easier to explore. The networks are constructed using collections of molecules with fixed stoichiometry (number of atoms) as nodes and stoichiometry-preserving transformation rules, i.e., bond breaks and bond formations, as edges. To generate the networks, a single node is initialized with a molecule or molecules and given the rule set defined, all possibilities are enumerated. The dimensionality of these networks can be calculated through the fractal dimension and shows that our networks are intrinsically lower dimensional compared to the full chemical space. Our representation can enable more efficient search strategies of chemical space by taking advantage of the lower dimensional nature of the networks.

Miko Stulajter, *Dmitrij Rappoport, and Filipp Furche*

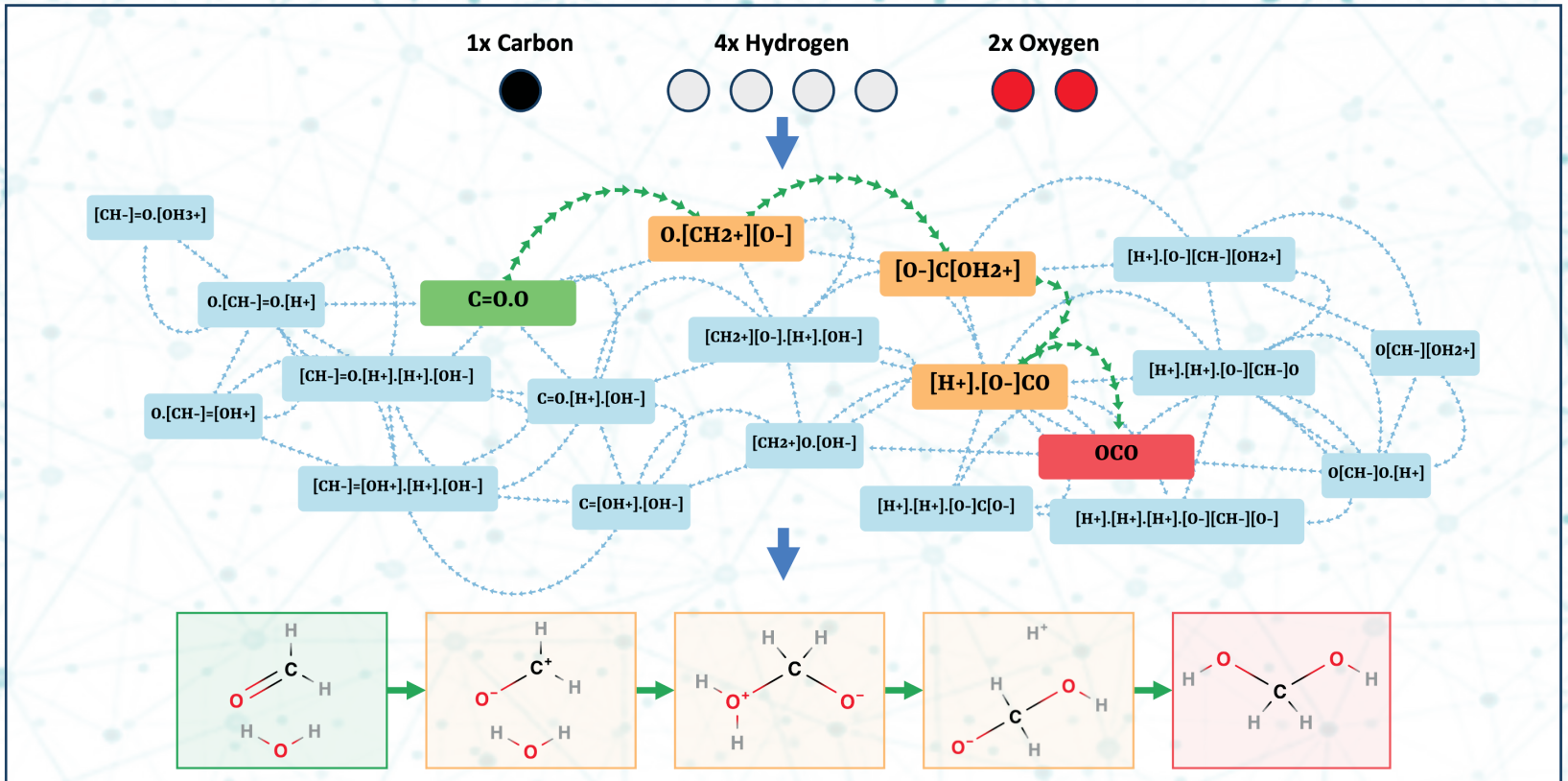
This research is supported by the National Science Foundation grants (CHE-2227112 and DUE-1930546) and the Computational Science Research Center (CSRC) at San Diego State University



Example network constructed with an initial molecule of hydrogen cyanide (HCN) composed of a carbon, nitrogen, and hydrogen atom (shown on top). The nodes are collections of molecules, and the edges are individual bonds being broken or formed.



Large network generated from hydrogen cyanide (HCN) and formaldehyde (CH_2O) composed of 5 atoms as shown on top. The path shown [orange nodes] rearranges the atoms from HCN and CH_2O [green node] to methyl isocyanate (CH_3NCO) [red node].



Network initialized from formaldehyde (CH_2O or C=O in SMILES notation) and water (H_2O or O). From top to bottom, we see the atoms used to generate the network, the network representation, and a diagram of the path with molecules. The network has node labels of SMILES strings which represent the molecules. The highlighted path [orange nodes] shows going from the reactants (C=O.O) [green node] to the product of methanediol ($CH_2(OH)_2$ or OCO) [red node]. The path diagram has molecule representations showing what bonds are broken and formed.