Molecular Discovery Strategies for Green Chemicals

Carles Estévez

Department of Green Chemistry. InKemia IUCT Group, Mollet del Vallès,. Spain – green-chem@iuct.com

Keywords: molecular discovery, diversity thresholds, green chemicals.

A remarkable feature of modern chemical technology is the diversity of substances needed to satisfy the wide variety of industrial uses. It is estimated that the number of chemicals needed to support our economy is around 70 000. Another characteristic of industrial chemicals is that a significant fraction display unintended environmental, health or safety (*EHS*) adverse effects. As a result, governments, industry and academia have launched a variety of environmental regulations and R&D programmes in order to reduce the use of hazardous chemicals.

Green Chemistry is the central cognitive framework aimed at expanding the chemical space with greener alternative chemicals. Discovery methods typically involve the screening of chemical libraries. One fundamental aspect of these innovation processes is whether the size and diversity of the search chemical spaces are sufficient to find new *EHS*-optimal chemicals with the desired efficacy of function.

The search of optimal molecules in chemical ensembles has been extensively investigated in the drug discovery field where medicinal chemists aim to find successful drugs through the *in silico* or biological screening of chemical libraries containing hundreds to millions of compounds. Empirical rules of the relationships between molecular diversity and the size of the drug candidate libraries have been obtained[1]. However, the fundamental laws and relationships between diversity, ensemble size, and property satisfiability remain unknown.

Consider a specific industrial function that needs to be fulfilled by a chemical substance. We can define the set of N_P functional variables that govern technical performance and the set of N_{EHS} properties that influence the *EHS* impact profile. Under a green chemistry approach, we will consider both sets of variables as essential design parameters. The sum $N_P + N_{EHS} = N$ defines a N-dimensional chemical space. We have theoretically investigated the satisfiability of such set of specifications in the context of a molecular discovery process.

Assuming a random search and negligible constraints between the N variables, the minimum size of the search space needed to find an optimal chemical depends on the number of parameters N and the mean fitness, f, of the search library (Fig. 1).

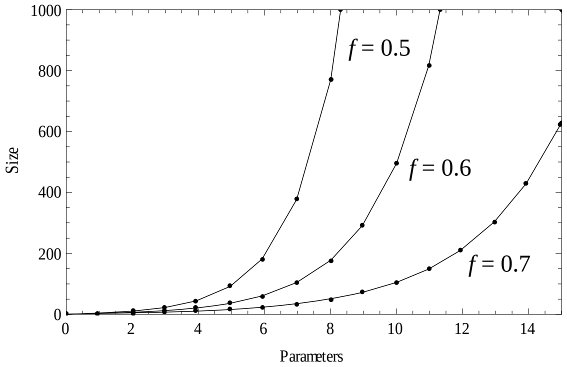

Fig. 1 – Influence of design dimensionality on the size of the chemical space necessary to find an optimal chemical (random search without parameter constraints). f is the mean fitness of the chemical search space.

The number of candidate chemical entities to be explored grows exponentially as the number of design parameters increases. A straightforward implication of this result for the design of green chemicals is that the pool of candidates to be explored has to be enlarged by orders of magnitude unless the fitness of the search ensemble is increased. A possible direction to achieve more fitted ensembles is to build *EHS*-optimal libraries from which molecular discovery processes centered on performance optimization could be conducted.

A critical aspect is whether these enlarged sets of chemicals have an adequate diversity. Molecular diversity is correlated with functional diversity since similar chemical structures generally lead to similar physico-chemical properties, functional performance and biological activity. In our analysis, focused on functional diversity, we note that the diversity of an ensemble is related to the number of interactions (influences) between the N design variables. Interactions arise whenever physico-chemical properties governing functional and/or *EHS* parameters are strongly (anti-)correlated. Fig. 2 shows a principal component analysis of molecular properties and *EHS*-variables of 250 organic solvents. Note the strong correlation between ecotoxicity and boiling point. On the other hand, the health hazard parameter does not seem influenced by other parametrs and its variation is relatively independent of the selected parameters. Such interactions may lead to conflicting constraints between variables that frustrate the identification of an optimal chemical in a molecular discovery process.

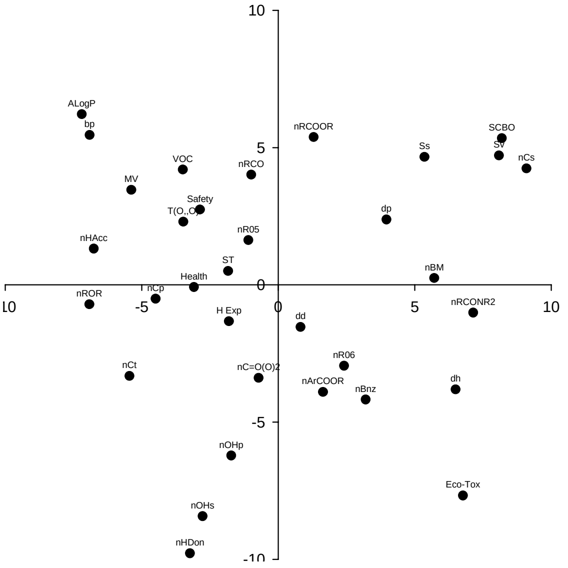

Fig. 2 – Principal components analysis of 33 solvent properties of the SOLVSAFE library. Certain *EHS* properties show strong correlations with many molecular features and properties.

We know that interactions effectively reduce the functional diversity of chemical ensembles (Fig. 3).

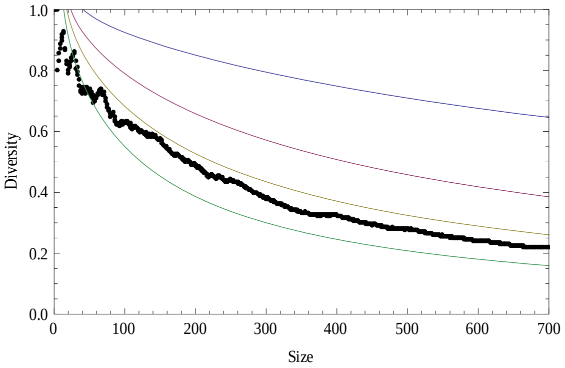

Fig. 3 – Molecular discovery dynamics showing the functional diversity as a function of ensemble size. Continuous lines are theoretical predictions. The upper line is a diversity line without interactions. The remaining theoretical lines below show different degrees of interactions. The dotted line is a numerical simulation with interactions between variables.

But is there any measurable diversity threshold below which the probability to find an optimal chemical is negligible? Our theoretical analysis, although preliminary, indicates that this is the case and numerical simulations of discovery processes with different degrees of interactions between design variables agree with the predicted thresholds.

We have applied the above analysis to the SOLVSAFE library[2], an ensemble composed of 249 *EHS*-optimized organic solvents for industrial applications. The experimental data indicates that its functional diversity is lower than the diversity expected for parameters free of interactions but very close to the critical diversity threshold. The fact that only one solvent out of 249 was found optimal seems to confirm the theoretical findings. More work is needed in this direction to fully validate the efficacy of molecular discovery strategies based on *EHS*-optimized ensembles.

Acknowledgements

The author thank Dr. Sergio Gómez for many helpful discussions on the mathematical treatment of diversity in the context of discovery assembled ensembles.

References

1 Y. C. Martin, *J. Comb. Chem*, 2001, **3** (3),231-250.

2 C. Estévez, *Green Solvents for Chemistry* in “Sustainable Solutions for Modern Economies”, ed. R. Höfer. RSC Green Chemistry Series, 2009.