A Computational Method for Design and Discovery of Optimal Molecular Scale Solar Antennas

Sofia Izmailov

1 Introduction

Molecular scale solar antennas are complex multi-component materials with many variables (e.g., structural features) for optimization. The number of potential materials is very large, and a complete search over a molecular library of candidates is unreasonable, as synthesis and property testing are expensive and time-consuming. This research focused on developing High Dimensional Model Representations (HDMR) [1] to estimate molecular property values using just a modestly-sized input of laboratory data. The method was tested on reaction yield data for the synthesis of trifluoromethylation performed by the MacMillan Group at Princeton.

2 Theoretical Background

Consider N independent variables describing each molecule in a given library. A physical property $F(\boldsymbol{x})$ (e.g., spectral absorbance) of a molecule in this library can be described as a function of these N independent variables: $\boldsymbol{x} \equiv (x_1, x_2, \ldots, x_i, \ldots x_N)$. For example, a library of 1,2,4-trisubstituted benzene molecules can have various properties described as a function of the three different functional groups around the benzene. These substitution sites would be the independent variables. A library where properties are defined in this way may be described with HDMR, ideally operating with a very modest sample of experimental data [1].

HDMR provides the mapping $\boldsymbol{x} \rightarrow F$ for functions $F(\boldsymbol{x}) = F(x_1, x_2, \ldots, x_i, \ldots x_N)$ with the specific goal of interpolating over $F(\boldsymbol{x})$ from a known coarsely sampled set of input values $\boldsymbol{x}^r, r = 1, 2, \ldots R$ and the associated outputs $F(\boldsymbol{x^r})$. In the case of a molecular library, this means that a small sample of the entire library needs to be synthesized and tested in order to create a map to estimate the properties of the untested molecules.

A function $F(\boldsymbol{x})$ is decomposed by HDMR into a sum of lower dimensional components[1]:

$$F(\boldsymbol{x}) = f_0 + \sum_{i=1}^{N} f_i(x_i) + \sum_{1 \le i < j \le N} f_{ij}(x_i, x_j) + \cdots + f_{12\ldots N}(x_1, x_2, \ldots, x_N) \qquad (1)$$

where each of the component functions represents the unique contribution of its variables to the value of the property $F(\boldsymbol{x})$: f_0 is the base contribution which is independent of the values of the $\boldsymbol{x}$ variables, $f_i(x_i)$ is the contribution of substituents at site i on a molecular scaffold, $f_{ij}(x_i, x_j)$ is the cooperative contribution of substituents at sites i and j, etc.

The HDMR expansion of $F(\boldsymbol{x})$ taken to the N-th order is exact [1]. In most realistic applications, HDMR component functions up to the second or third order are sufficient to quantitatively describe the input-output relationships $\boldsymbol{x} \rightarrow F$[2]. Consequently, it is expected that $F(\boldsymbol{x}) \approx G(\boldsymbol{x}) = f_0 + \sum_{i=1}^{N} f_i(x_i) + \sum_{1 \leq i < j \leq N} f_{ij}(x_i, x_j)$ should be adequate for representing molecular properties; this coincides with the statement that up to pairwise cooperative interactions between substituents in a molecule are expected to sufficiently capture the $\boldsymbol{x} \rightarrow F$ relationship. In this fashion, HDMR reduces the initial function of N variables to a set of low dimensional component functions that can be systematically deduced from laboratory data [1, 2]. In this work the truncated HDMR decomposition was used for molecular property prediction.

3 Method and Results

In order to develop the HDMR-based method, we used actual laboratory data as a test. In particular, HDMR was used to estimate reaction yields, the "property" of the reaction shown in Figure 1. Although reaction yield is not directly a property, the mathematical and physical nature of the problem is fully parallel to what will be encountered in the discovery of effective molecular antennas. There were seven variables (i.e, chemical species, and solvents) that determine the reaction yield. Upon discretization of these reaction variables, a total of 10^6 possible experimental setups could be made. In practice the synthesis experiment was carried out for a modest sampling of 10^3 setups. This sampling was guided by a stochastic algorithm[3].

We deduced the HDMR component functions using the laboratory data for observed yields which were at least 1% (641 distinct reaction setups). The HDMR estimation quality was evaluated using cross-validation. We set aside one laboratory-tested reaction setup as a "test case" and used the other 640 setups as the "training set" to generate the HDMR map and estimate the reaction yield of the test sample. This procedure was repeated for all of setups tested in the laboratory to get their estimated reaction yields. The estimated reaction yields of the data were used to assess the quality of the HDMR by its ability to accurately estimate the yield values.

The truth plot in Figure 2 of all 641 samples shows the estimation quality of the reaction yields. Considering the small amount of the data, the overall quality is excellent. The demands here do not require absolute prediction of the single best compound out of the 10^6, but rather the HDMR serves as a guide for further reaction setups which are more likely to give high yield.

4 Future Work

The same logic in the test case above can be applied to general molecular and material properties, including candidates for efficient solar energy collection. In particular, if a library of molecular scale solar antennas or any other materials of interest can be represented as a function of multiple independent variables, then the HDMR-based estimation procedure can be used to guide automated synthesis towards molecules and materials which have desired property values. This approach should decrease the search time and syntheses needed to

find the target materials with desired properties.

References

[1] Genyuan Li, Carey Rosenthal, and Herschel Rabitz. High dimensional model representation. *J. Phys. Chem. A*, 105(33):7765–7777, 2001.

[2] Genyuan Li, Sheng-Wei Wang, and Herschel Rabitz. Practical Approaches to Construct RS-HDMR Component Functions. *J. Phys. Chem. A*, 106:8721–8733, 2002.

[3] Information from Ofer Shir MacMillan Group. Princeton University.

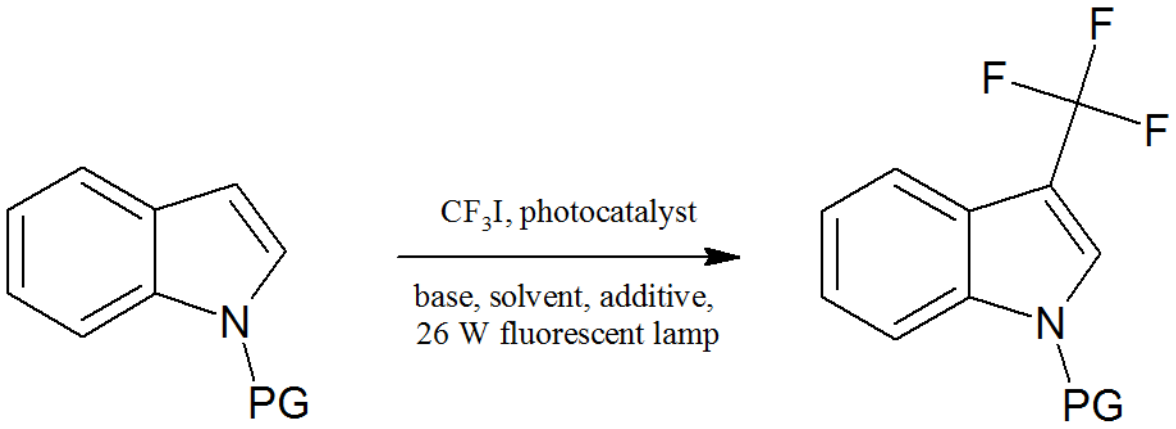

Figure 1: The general scheme of the fluoromethylation reaction considered to test the effectiveness of HDMR when working with laboratory data. PG is the protecting group on the nitrogen atom. The reaction yield depends on a set of seven laboratory variables. These variables were the protecting group, photocatalyst, base, solvent, additive, amount of solvent, and amount of CF_3I.

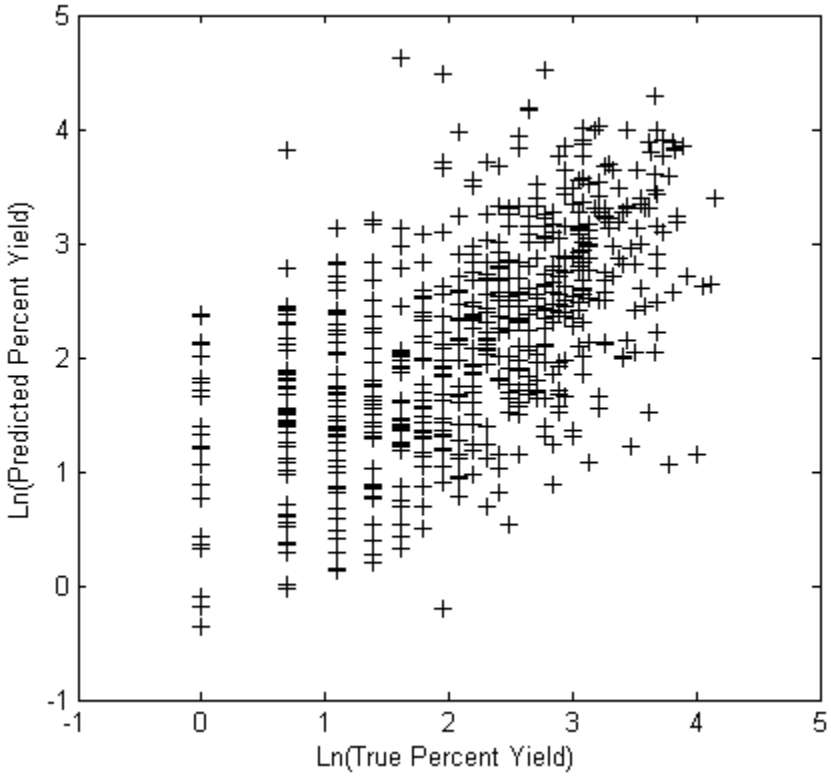

Figure 2: Truth plot for percent yield. A total of 641 reaction setups were used in a cross-validation evaluation of HDMR. The results are of quite acceptable quality to guide additional synthesis experiments.