Chemical Physics Letters 572 (2013) 1–12

Contents lists available at SciVerse ScienceDirect

Chemical Physics Letters

journal homepage: www.elsevier.com/locate/cplett

FRONTIERS ARTICLE

Discovering predictive rules of chemistry from property landscapes

Katharine W. Moore Tibbetts [1], Richard Li, István Pelczer, Herschel Rabitz *

Department of Chemistry, Princeton University, Princeton, NJ 08544, United States

ARTICLE INFO

Article history:
Available online 27 March 2013

ABSTRACT

Predicting the chemical and physical properties of molecules often relies on systematic rules relating the properties to molecular characteristics. This Letter introduces a novel method to reveal predictive chemical rules based on analysis of the *chemical property landscape*, which specifies the functional relationship between a measured property and an appropriate set of molecular variables. As an illustration, we consider landscapes relating the ^{17}O NMR chemical shift, ^{13}C NMR chemical shift, and IR vibrational frequency to the moieties attached to a carbonyl group. Implications of this 'Chemscape' formulation for general molecular property prediction are discussed.

1. Introduction

Systematic rules relating physical or chemical properties to molecular characteristics commonly form the basis for predicting the properties of new compounds. For example, the energy gap between the filled and empty orbitals in organometallic complexes may be predicted by the ligand spectrochemical series [1], and the Hammett equation predicts the ionization rate of substituted benzoic acids with respect to the moieties on the phenyl group [2]. Such rules based on an appropriate descriptor of molecular structure may often be understood in terms of the chemical nature of the ligands or moieties.

This work presents a novel procedure to identify systematic rules for property prediction that rests on analysis of the *chemical property landscape*, which is the functional relationship between the property value $\mathcal{P}$ and a set of n variables v_i [3]

$$\mathcal{P} = f(v_1, \ldots v_n). \quad (1)$$

The creation of a chemical property landscape starts with experimental observations of a desired property over a family of related molecules. Subsequently, a choice must be made for an appropriate set of variables to describe the molecular family. For instance, commonly employed regression methods such as Quantitative Structure Activity Relationship (QSAR) typically rely on $n \sim 10 - 100$ or more chemoinformatic descriptors acting as variables in Eq. (1) [4–6], which can make it difficult to distill simply understood rules from the resulting high-dimensional property landscapes. In contrast, this paper will show that a special small set of variables can greatly simplify the analysis of predictive rules in terms of the chemical characteristics of the molecules.

Recently, we demonstrated that the topology (i.e., local and global maxima and minima) of chemical property landscapes may be assessed generically [7,8]. The latter analysis, called 'OptiChem' theory, rests on two key Assumptions: (i) the property is physically well-defined, and (ii) the chosen variables adequately permit free movement on the landscape (see Refs. [7,8] for details). Most physical or chemical properties should satisfy Assumption (i), but exceptions can arise (e.g., the ratio of two well-defined properties to specify a new 'property' will likely be outside the validity of Assumption (i) due to the restricted form of physical observables). It is difficult to assess whether Assumption (ii) is satisfied *a priori* with a specific set of variables. However, upon satisfaction of both Assumptions, mathematical analysis [7,8] shows that the topology of any chemical property landscape should be *monotonic*, i.e., the landscape contains extrema only at globally maximal and/or minimal property values (as well as possibly saddles at intermediate property values). These topological conclusions arise naturally from either a classical or quantum–mechanical description of an 'open' system interacting with its environment [9]. Landscape topology is illustrated in Figure 1 with two variables $v_1 = \mathrm{X}$ and $v_2 = \mathrm{Y}$ in Eq. (1), where the landscape (a) contains three maxima with different property values and the landscape (b) is monotonic. Employing an insufficient number and/or type of variables may violate Assumption (ii) of OptiChem theory and cause local extrema to arise on the landscape, as in Figure 1a.

Extensive evidence from the chemical literature suggests that often only a few *well-chosen* variables are needed to satisfy Assumption (ii) of OptiChem theory for diverse properties including catalytic activity and inhibition of a protein, as the reported landscapes are monotonic [7,8]. We also recently demonstrated the existence of monotonic property landscapes relating the ^{13}C

* Corresponding author. Fax: +1 609 258 0967.
E-mail address: hrabitz@princeton.edu (H. Rabitz).
[1] Current address: Department of Chemistry, Temple University, Philadelphia, PA 19122, United States.

0009-2614/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.cplett.2013.03.040

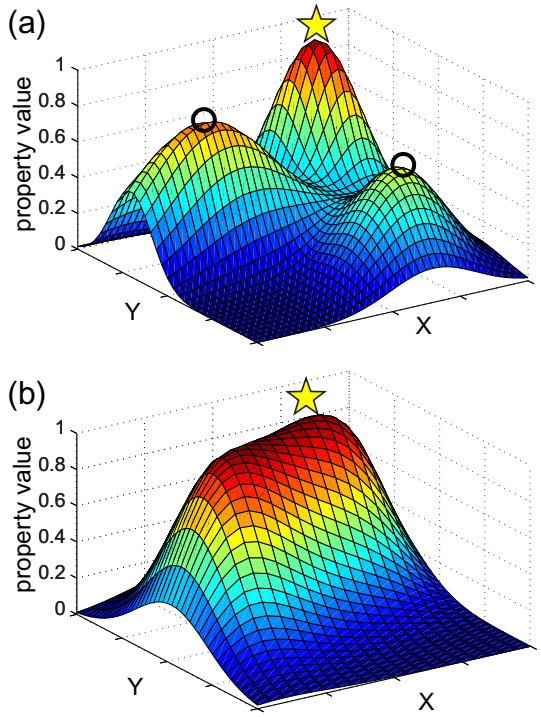

Figure 1. Schematic illustration of chemical property landscapes in Eq. (1) as a function of two variables $v_1 = X$ and $v_2 = Y$. In landscape (a), there are three distinct maxima (hills) with different property values. The global maximum is marked with a yellow star, and the local suboptimal maxima are marked with a black circle ○. Chemical property landscapes should have a monotonic topology depicted in (b) upon satisfaction of the two Assumptions of OptiChem theory [7,8].

NMR chemical shift to moieties on molecular scaffolds, where the property landscapes revealed rules that enabled highly accurate ^{13}C NMR chemical shift prediction of untested compounds [10]. The present work extends the latter results to consider not only (a) identifying a monotonic landscape and the associated rules for a desired chemical property, but also (b) the additional goal of explaining the resulting rules in terms of the chemical characteristics of the variables.

The goals (a) and (b) comprise what will be referred to as 'Chemscape', a methodology that takes advantage of the principles of OptiChem theory to facilitate chemical property prediction and analysis. Chemscape is illustrated in this work for landscapes relating spectral properties of a family of molecules sharing a common scaffold and sets of chemically bonded moieties. In particular, we will illustrate Chemscape for a scaffold with two sites for substituent bonding and thus two sets of variables $v_1 = \mathrm{X}$ and $v_2 = \mathrm{Y}$ in Eq. (1). We will show that this scenario can produce rules for goal (a), and that chemically motivated principles can then be applied to explain the origin of the rules for goal (b). The chemical names, expressed as integer labels, of the moieties at each site form a complete set of variable *values* (e.g., the values of X and Y, shown later). By extension, a molecular family with N sites on a scaffold can be described by only the minimal set of N enumerated moiety variables [11,10], as described in Section 2.

Although this work uses spectral properties to investigate goals (a) and (b), Chemscape can be applied to *any* common chemical or physical property of a family of related compounds upon an appropriate definition of the variables, provided that they satisfy Assumption (ii) of OptiChem theory. In cases where more than two variables are needed, the resulting high-dimensional property landscapes may be treated using special analysis methods in order to extract the corresponding predictive rules [11] (i.e., achieve goal (a) above). Once again, subsequent analysis of these rules in terms of the chemical characteristics of the variables would aim to accomplish goal (b).

The remainder of the paper is structured as follows. Section 2 describes the methods used to reveal whether monotonic property landscapes exist and specifies the meaning of a property prediction 'rule' in this context. Section 3 presents NMR and IR spectral property landscapes and explains the resulting predictive rules in terms of the electronic influences of the moieties. The rules revealed by Chemscape are shown to facilitate spectral property prediction for new compounds, as well as enable the analysis of ^{13}C NMR and IR spectra *simultaneously*. Finally, Section 4 discusses the wider implications of Chemscape and presents concluding remarks.

2. Chemscape methodology

OptiChem theory predicts that chemical property landscapes should be both smoothly varying and monotonic upon satisfaction of Assumptions (i) and (ii) above [7,8]. Because chemical moieties constitute discrete variables, it is not evident *a priori* that the corresponding landscapes can be smoothly varying (albeit with inherent finite resolution) and monotonic. Here, we simply choose chemical moiety labels to act as the variable values, as recently introduced [8,10–14]; for example, the moieties —CH_3, —OH, —Cl, ... considered for bonding at a particular site may be arbitrarily given integer labels 1, 2, 3, ..., respectively. Thus, the appearance of the landscape depends on the 'ordering' $\mathcal{O}$ of the moiety labels along the axes of the landscape (i.e., with each axis corresponding to the associated site for moiety bonding on the molecular scaffold). In some cases, smooth chemical property landscapes have been revealed by choosing the moiety ordering $\mathcal{O}$ based on chemoinformatic descriptors such as Hammett constants [15] or the 'flexibility index' in a polymer [16]. However, if the contribution of each moiety to the property value is not known *a priori* and/or large numbers of moieties are involved, identifying an ordering $\mathcal{O}_s$ that generates a smooth property landscape appears to present a difficult task. For N_i moieties at the ith scaffold site, there are $N_i!$ unique moiety orderings that may be sampled, and the total number of possible orderings for a family of molecules is $\prod_i N_i!$ considering all sites.

To accomplish goal (a) of Chemscape, the first step is to assess whether a monotonic property landscape as a function of the variables exists, upon identification of an appropriate ordering $\mathcal{O}_s$. If a monotonic landscape exists, then the ordering $\mathcal{O}_s$ defines a 'rule', i.e., *relationship between the property and the moiety ordering*. The ordering $\mathcal{O}_s$ may be used to predict property values of compounds for which no experimental data are available even without further chemically based characterization of the moieties involved. Such ordering rules were found to enable accurate prediction of ^{13}C NMR chemical shifts in several families of organic compounds through interpolation over the landscape [10]. In general, multiple orderings $\{\mathcal{O}_s^{(r)}\}$, each corresponding to a distinct rule $(r), r = 1, 2, \ldots$, may produce an acceptable (i.e., smooth and monotonic) property landscape over the experimental data. To accomplish goal (b) (i.e., explain the origins of a rule based on the chemical characteristics of the moieties), it is desirable to obtain a 'chemically motivated' ordering $\mathcal{O}_{\mathrm{chem}}$ from among the possibilities $\{\mathcal{O}_s^{(r)}\}$ based on known contributions of the moieties to the desired property, if they are available. When the chemical origins of the moiety contributions are unknown *a priori*, detailed analysis of a collection of orderings $\{\mathcal{O}_s^{(r)}\}$ may be necessary to achieve goal (b). This analysis may also offer the opportunity to identify the particular chemical features of the moieties that contribute. For the present work, the extensive chemical literature on moiety contributions to