RESEARCH ARTICLE | JULY 05 2022

Building machine learning assisted phase diagrams: Three chemically relevant examples

Xabier Telleria-Allika; Jose M. Mercero; Xabier Lopez; ... et. al

AIP Advances 12, 075206 (2022)
https://doi.org/10.1063/5.0088784

Building machine learning assisted phase diagrams: Three chemically relevant examples

Cite as: AIP Advances 12, 075206 (2022); doi: 10.1063/5.0088784
Submitted: 17 April 2022 • Accepted: 9 June 2022 •
Published Online: 5 July 2022

Xabier Telleria-Allika, Jose M. Mercero, Xabier Lopez, and Jon M. Matxain[a)]

AFFILIATIONS

Kimika Teorikoa, Polimero eta Material Aurreratuak: Fisika, Kimika eta Teknologia Saila, Kimika Fakultatea, Euskal Herriko Unibertsitatea UPV/EHU and Donostia International Physics Center (DIPC), P.K. 1072, 20080 Donostia, Euskadi, Spain

[a)]Author to whom correspondence should be addressed: jonmattin.matxain@ehu.eus

ABSTRACT

In this work, we present a systematic procedure to build phase diagrams for chemically relevant properties by the use of a semi-supervised machine learning technique called uncertainty sampling. Concretely, we focus on ground state spin multiplicity and chemical bonding properties. As a first step, we have obtained single-eutectic-point-containing solid–liquid systems that have been suitable for contrasting the validity of this approach. Once this was settled, on the one hand, we built magnetic phase diagrams for several Hooke atoms containing a few electrons (4 and 6) trapped in spheroidal harmonic potentials. Changing the parameters of the confinement potential, such as curvature and anisotropy, and interelectronic interaction strength, we have been able to obtain and rationalize magnetic phase transitions flipping the ground state spin multiplicity from singlet (nonmagnetic) to triplet (magnetic) states. On the other hand, Bader's analysis is performed upon helium dimers confined by spherical harmonic potentials. Covalency is studied using descriptors as the sign for $\Delta\rho(r_C)$ and $H(r_C)$, and the dependency on the degrees of freedom of the system is studied, i.e., potential curvature ω^2 and interatomic distance R. As a result, we have observed that there may exist a covalent bond between He atoms for short enough distances and strong enough confinement. This machine learning procedure could, in principle, be applied to the study of other chemically relevant properties involving phase diagrams, saving a lot of computational resources.

I. INTRODUCTION

Phase diagrams are graphical representations of functions f such that, given a state vector $\mathbf{x} \in \mathbb{R}^n$, they tell us the phase C of such state being the latter a categorical variable ($f : \mathbf{x} \in \mathbb{R}^n \rightarrow C$). These diagrams are extensively used in physical, chemical, and material sciences to understand how the properties of macroscopic systems change with external parameters, such as pressure, temperature, composition, etc. Hence, it is interesting to have a procedure to obtain the graphical representations of such f functions.

The brute force approach to obtain the phase diagrams goes as follows: the parameter space is uniformly discretized such that a grid composed of state points is obtained. Upon these state points, experiments or simulations can be carried out to determine the phase (label) of each point, hence, obtaining a phase diagram of the desired region. Even though this approach is conceptually simple and straightforward to automatize, it also carries three main drawbacks. First, since all points in the grid must be labeled, the process is expensive in terms of time and resources. Second, the lack of efficiency since there is no feedback coming from already obtained results. Finally, there are cases in which assigning a phase to a given point in the phase diagram is not possible and alternative methods must be employed.[1]

As it has also been done in many fields of chemical sciences,[2–14] in order to improve the efficiency of sampling the discretized parameter space, machine learning techniques have been employed such that phase boundaries have been accurately described while evaluating fewer points than in the straightforward approach.[15,16]

In this work, we have focused our attention to some specific machine learning techniques, i.e., semi-supervised learning and uncertainty sampling, and have used first principle calculations in order to label each sampled point. Contrary to previous studies, instead of applying the Label Propagation (LP) algorithm,[17] we employ Random Forests (RF) to compute the required uncertainties. The main advantage of this approach is that, contrary to LP,

neither a large inverse matrix nor an iterated Markov chain process should be computed. Hence, information can be propagated to denser grids in a faster way. As a benchmark, we have tested the efficiency of this method on ideal solid–liquid systems that contain a single eutectic point and a total number of four phases appear in the neighborhood of such point. In order to extend this procedure to other chemical properties involving phase diagrams, we have used two simple systems: a few electron Hooke atoms for studying ground state magnetic properties and He dimers for studying covalency.

For the former, we build ground state spin multiplicity phase diagrams for 4 and 6 electrons in a ellipsoidal Hooke atom taking as parameters the confinement potential curvatures and the electron–electron interaction screening parameter $(\omega_x^2, \omega_y^2, \omega_z^2, \lambda)$ in $\mathbb{R}_+^4$ expressed as in Hamiltonian 1. As far as noble gas molecules are concerned, previous computational works in the field of such dimers confined in nanoclusters and nanotubes have reported internuclear distances in the range of $[1.265, 2.447]\mathring{A}$ for He_2 dimers.[18–26] All these works report shorter interatomic distances between the noble gas atoms. For some of these systems, bond descriptors as in Bader's theory, such as Laplacian of the density at the Bond Critical Point (BCP) $\Delta\rho(r_C)$ and total energy density at the BCP $H(r_C)$, are negative; this insinuates there may be covalent bonds between noble gas atoms under these conditions.[22–24,26] In the present work, we study how covalency descriptors evaluated at the BCP evolve by varying the spherical confinement strength ω^2 as well as the internuclear distance R, both parameters contained in Hamiltonian 2,

$$H = \sum_{i=1}^{N}\left(-\frac{1}{2}\nabla_{\mathbf{r}_i}^2 + \frac{1}{2}\mathbf{r}_i^{\top}\mathbf{W}\mathbf{r}_i\right) + \sum_{j\neq i=1}^{N}\frac{e^{-\lambda r_{ij}}}{r_{ij}},$$
$$\mathbf{W} = \begin{pmatrix} \omega_x^2 & 0 & 0 \\ 0 & \omega_y^2 & 0 \\ 0 & 0 & \omega_z^2 \end{pmatrix}, \tag{1}$$

$$H = -\frac{1}{2}\sum_{i=1}^{4}\nabla_{\mathbf{r}_i}^2 - 2\sum_{J=1}^{2}\sum_{i=1}^{4}\frac{1}{r_{iJ}} + \sum_{j\neq i=1}^{4}\frac{1}{r_{ij}} + \frac{1}{2}\sum_{i=1}^{4}\mathbf{r}_i^{T}\mathbf{W}\mathbf{r}_i + \frac{1}{R}. \tag{2}$$

The aim of this work has, therefore, been to set up a procedure in which chemically relevant properties, such as ground state spin multiplicity and covalency in chemical bonding, change character induced by external confinement potentials. This procedure could be systematically generalized to the study of other chemically relevant properties involving any type of phase diagram. The obtained results indicate that this process enables to obtain sharper boundaries between the phases without the need of using any hyperparameter. All of these, without losing the efficiency of uncertainty sampling, which, in our case, has enabled us to build proper phase diagrams using less than 2% of the points with respect to naive grid computations.

II. COMPUTATIONAL METHODS

A. Thermodynamic model for eutectic composition

As it is known, when two substances are mixed together their thermodynamical properties, such as vapor pressure, melting and boiling temperatures, specific volume, etc., change depending on the activities of both substances. A mixture is considered to be ideal if the interactions between the molecules of both substances interact among them with similar interactions (in nature and strength) such that the difference in potential energies effects can be neglected. As a matter of fact, for ideal mixtures, the activities of each component correspond directly to their molar fraction.

We shall now consider a two component liquid–solid system at constant pressure, which has the following properties and approximations:

- Both components are ideally miscible in the liquid phase.
- Both components are ideally immiscible in the solid phase (they form a heterogeneous solid).
- Fusion enthalpies do not depend strongly on temperature.

Bearing these properties in mind, one shall derive a relation that determines the molar fraction of each component x_i as a function of their respective fusion temperatures T_{f_i}, fusion enthalpy $\Delta H_{f,i}$, and system temperature T for each component as in the following equation:

$$\ln x_i = \frac{\Delta H_{f,i}}{R}\left(\frac{1}{T_{f,i}} - \frac{1}{T}\right), \sum_{i=1}^{2} x_i = 1. \tag{3}$$

The nonlinear system 3 can be solved using previously studied methods[27] such that the eutectic composition vector $\mathbf{x}_{eu}$ and the eutectic temperature T_{eu} are obtained. Now, for two component systems, once we obtain (x_{eu}, T_{eu}), we shall label any point in the (x_1, T) plane using the eutectic point coordinates as well as the melting curves given by Eq. (3), as shown in Fig. 1.

Thus, by using this rather simple scheme, we have been able to test the uncertainty sampling method on the systems that are of similar nature as the ones reported in previous studies[15,16] and found out that it is suitable.

B. Quantum chemical methods

1. Hooke atoms

Hooke atoms can be described by a general one body Hamiltonian that reads as in (1). Even if there are many works in the bibliography that study electronic structure problems based on Hamiltonians as (1),[28] nowadays it is still an interesting topic that can be

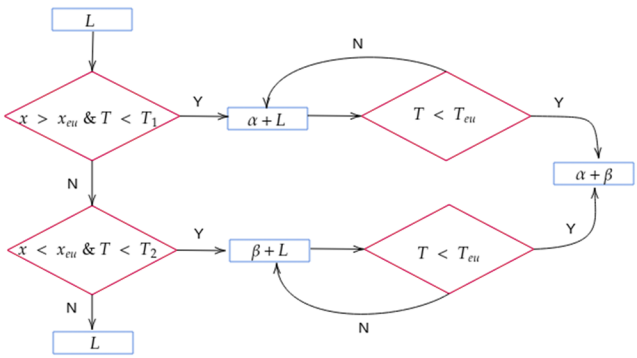

FIG. 1. Decision tree used for labeling each point in the (x_1, T) plane using eutectic point coordinates and melting curves.

related to confined quantum systems,[29,30] electron correlation,[31,32] and many body theory.[33,34] As far as the symmetry of the Hamiltonian (2) is concerned, the diagonalizable matrix **W** contains information of the external (confinement) potential; depending on the degeneracy of this matrix g, the system may be spherical ($g = 3$), spheroidal ($g = 2$), or tri-axial ellipsoidal ($g = 1$).

In this work, we have studied the case where the matrix **W** is diagonal with eigenvalues ω_q^2 that may have three- or twofold degeneracy and the two body interaction is driven by a screened Coulomb potential where $\lambda \geq 0$. The necessary two body integrals were calculated and coded in GAMESS US in previous studies realized in our team.[35] From a previous study on spherical Hookean atoms,[36] we have been able to obtain some optimized even tempered basis sets for describing systems containing even number electrons ($n = 2, 4, 6, 8, 10$) with curvature $\omega^2 = 0.25$ at MRPT2(n,13) level; after careful inspection, we reached the conclusion that the best compromise between size and accuracy was given by the basis set obtained for 6 electron system in the singlet spin state, namely, ETBS-6S. Based on this result, in this work, we have used these basis functions; nevertheless, we have decided to reduce the active space down to 10 active orbitals. By doing so, electronic structure computations are faster and the error for $\omega^2 = 0.25$ with respect to the large active space is in the order of 1×10^{-3} hartree for absolute energies while as for the singlet–triplet gap is of the order of 8×10^{-2} eV. Either way, for larger or smaller active spaces, the absolute error with respect to reference values is of the order of 1×10^{-2} hartree, while the singlet–triplet energy gap is of the order of 5×10^{-2} eV. Energies for both active spaces and reference values are summarized in Table I.

2. He dimers

Several authors have studied confined chemically interesting systems as atoms and molecules by means of harmonic potentials.[40] In this study, we have considered systems composed by two He atoms ($Z = 2$) with net null charge and singlet spin state. In each case, the nuclei are treated as in the Born–Oppenheimer approximation and they are centered in positions $\mathbf{R}_1$ and $\mathbf{R}_2$ with respect to the origin. Besides, we include a quadratic potential to model the confinement effects; this potential is represented by a diagonal quadratic form **W** with elements ω^2. The electronic Hamiltonian for such system can be written in atomic units as in Eq. (2), where lower case letters refer to electrons and capital letters refer to nuclei. As a shortcut, we may name $R = |\mathbf{R}_1 - \mathbf{R}_2|$ as the internuclear distance since this is the most chemically relevant coordinate. In a first approximation, we have computed ground state singlet ($S = 0$) Hartree–Fock wave functions augmented correlation consistent triple zeta basis functions aug-cc-pVTZ for spherical harmonic potentials ($\omega_x^2 = \omega_y^2 = \omega_z^2$) in a given point in the feature space (ω^2, R). Zooming into areas of interest, we have also performed CASSCF(4,10)/6-31++G** calculations in order to gain a deeper understanding of these systems. Electronic structure calculations have been carried out using GAMESS-US software. Once we have obtained the wavefunction, we perform Bader's analysis upon the electronic density using AIMPAC software. The main features we extract from this analysis are the Laplacian of the electron density at the bond critical point $\Delta\rho(r_C)$, kinetic energy density at the bond critical point $G(r_C)$, potential energy density at the bond critical point $V(r_C)$, and total energy density at the bond critical point $H(r_C)$.

C. Semi-supervised learning techniques

Semi-supervised learning techniques are halfway between supervised and unsupervised techniques and are employed to obtain information connecting the feature set $X = \{x_1, \ldots, x_n\}$ with the label set $Y = \{y_1, \ldots, y_n\}$ despite the small number of available data.[41–43] In the context of phase diagram creation, once the parameter space has been discretized as a grid X, we may split these grid points into two set: a labeled set $\mathscr{L}$ (for which the phase is known) and an unlabeled set $\mathscr{U}$ (for which the phase is not known). Namely, the labeled set $\mathscr{L}$ is composed by points formed as an ordered Cartesian products $(x_1, y_1), \ldots, (x_l, y_l)$, where $Y_L = \{y_1, \ldots, y_l\} \in \{1, \ldots, C\}$ is the set of possible labels (categories). Then, we also have the unlabeled set $\mathscr{U}$ composed by $(x_{l+1}, y_{l+1}), \ldots, (x_{l+u}, y_{l+u})$ points where the labels $Y_U = \{y_{l+1}, \ldots, y_{l+u}\} \in \{1, \ldots, C\}$ are unknown and the cardinality of this set is much larger than the former one, $l \ll u$. By doing so, we shall know where to map the next experiment to obtain phase diagrams as it has been done in very recent studies.[15,16]

Zu and Lafferty developed an algorithm named "Label Propagation Algorithm"[17] in which information from labeled data is propagated to unlabeled data in a stochastic process by means of Markov chains. Here, one defines a graph containing all data—labeled and unlabeled—and defines the connection strength (weight) by pairs as in (4). We may observe the closer two points are, the stronger they interact and the interaction also depends on a hyper-parameter σ that can be tuned. As it was proved in Ref. 17, the Markov process 5 has a unique solution regardless the initial condition for **Y** and is given by 6,

$$w_{ij} = \exp\left(-\frac{\sum_{d=1}^{D}(x_i^{(d)} - x_j^{(d)})^2}{\sigma^2}\right), \tag{4}$$

TABLE I. Energies in atomic units and energy gaps Δ_{T-S} in eV for singlet and triplet states for spherical Hooke atoms composed by 4 and 6 electrons with $\lambda = 0.0$ and $\omega^2 = 0.25$.[28,37–39]

	MRPT2(n,10)/ETBS-6S			MRPT2(n,13)/ETBS-6S			References		
n	Singlet	Triplet	Δ_{T-S}	Singlet	Triplet	Δ_{T-S}	Singlet	Triplet	Δ_{T-S}
4	6.3912	6.3529	−1.04	6.39010	6.3520	−1.04	6.3855	6.3488	−1.00
6	12.0846	12.0479	−1.00	12.0828	12.0461	−1.00	12.0663	12.0313	−0.95

$$\begin{bmatrix} \mathbf{Y}_l \\ \mathbf{Y}_u \end{bmatrix}^{(n+1)} = \begin{bmatrix} \tilde{\mathbf{T}}_{ll} & \tilde{\mathbf{T}}_{lu} \\ \tilde{\mathbf{T}}_{ul} & \tilde{\mathbf{T}}_{uu} \end{bmatrix} \begin{bmatrix} \mathbf{Y}_l \\ \mathbf{Y}_u \end{bmatrix}^{(n)}, \tag{5}$$

$$\mathbf{Y}_u = (\mathbf{I} - \tilde{\mathbf{T}}_{uu})^{-1} \tilde{\mathbf{T}}_{ul} \mathbf{Y}_l. \tag{6}$$

The main drawback of this method is the fact that, for dense or large enough grids, the computation of the matrix $(\mathbf{I} - \tilde{\mathbf{T}}_{uu})^{-1}$ is computationally expensive or, in the worst case, the matrix is singular. This problem may be sidestepped by, instead of using the closed formula (6), taking an initial label vector and iterating it as in Eq. (5). Nevertheless, since convergence is reached after several iterations and large matrices are still involved, this algorithm is still costly for dense enough grids and this deficiency is more remarkable as the number of dimensions of the feature space increases. As an alternative, other, rather cheap, classifiers, such as Random Forests (RF), can be used to solve the problem. Since RF do not have over fitting problems, using a large enough number of trees, the method is universal for all systems studied in this work. Therefore, the initialization set is employed to train a RF model that will label the unlabeled points assigning them to each category with a given probability, that is, an equivalent approach as using label propagation.

Once information is propagated to all unlabeled points, we must compute the labeling uncertainty in order to select the next point to be computed. Since we obtain the probability for each point $\mathbf{x} \in \mathscr{U}$ to belong to each phase $P(C|\mathbf{x})$ using RF, we compute the uncertainty of such point $u(\mathbf{x})$ using Shannon entropy[44] as in Eq. (7) where the sum is performed upon all categories (phases),

$$u(\mathbf{x}) = -\sum_C P(C|\mathbf{x}) \log P(C|\mathbf{x}). \tag{7}$$

D. Work procedure and calculation setup

In this work, we have been following some steps repeatedly to obtain our solid–liquid, magnetic, and covalency phase diagrams; let us first enumerate them and give a further explanation about them (see also Fig. 2):

1. Create a suitable grid in the parameter space so that all categories will, in principle, be represented.
2. Select some initialization points in the grid.
3. Carry out sorting tree or required electronic structure calculations upon the initialization points and assign them a label corresponding to the output of the calculations.
4. Propagate the information from already computed grid points using trained RF model to the rest to obtain the guessed labels.
5. Compute Shannon's entropy and select the maximum entropy point to carry the computations in step 3 and repeat until maximum number of calculations is reached.

Step 1: We need all classes phases (in our case, all four phases concerning the solid–liquid system, at least a ground state singlet and a ground state triplet or a covalent and non-covalent bond) in order to start the sampling procedure; thus, we must build a grid in the parameter space of each system suitable for this. For the solid–liquid diagram, the composition x is in the range [0,1] and the temperature T is in the range $[0.25 \times \min(T_1, T_2), 1.25 \times \max(T_1, T_2)]$ with a total number of 50 000 grid points. In the case of Hooke atoms, the confinement strength parameters $\omega^2_{x,y}, \omega^2_z$ have been taken in the range $[0.20, 0.30]$ since optimized basis set for spherical systems with $\omega^2 = 0.25$ was available[36] and the screening parameter λ was taken in the range [0.05, 2.55] with the aim of representing slightly and strongly screened systems accounting for a total number of 3375 grid points. Finally, in relation to He dimers, broad ranges in interatomic distance $R \in [1.0, 4.0]$ Å and confinement strength $\omega^2 \in [0.0, 4.0]$ were employed, giving rise to a total number of 250 000 grid points.

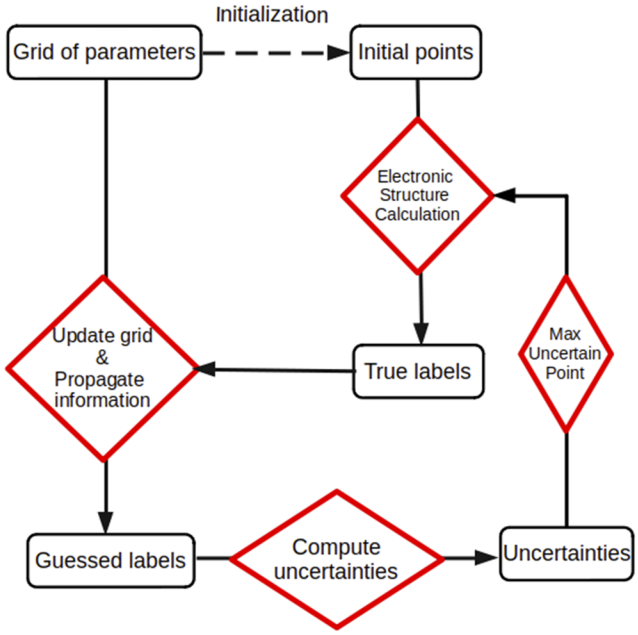

FIG. 2. Flowchart for the general procedure for phase diagram construction.

Step 2: We select the initialization points in the grid built in step 1. For the solid–liquid system, a total number of 50 uniform distributed points where selected taking the Cartesian product of 5 equidistant points for x and 10 equidistant points for T. With respect to Hooke atoms, by hypothesis, the edges of the parameter grid would contain all different phases; therefore, since three parameters were employed, the initialization set was obtained using the extreme 8 points. Meanwhile, for the He dimers, the extrema of the grid were considered along with 100 points selected at random.

Step 3: Once a given point in the parameter grid is selected (either the initialization ones or a new one coming from step 5), we carry out sorting tree or required electronic structure calculations in GAMESS US and we assign a label to this point depending on the result obtained by the sorting tree, the sign of the energy gap Δ_{T-S}, or density Laplacian or total energy density.

Step 4: We propagate the label information from the data points to the unlabeled ones using 200 trees to form the RF.

Step 5: With the information from the previous step, we compute Shannon's entropy to all guessed points in the grid and select

the one with the highest score. We take this point out of the unlabeled points pool $\mathscr{U}$, perform electronic structure calculations as in step 3, and update the labeled database $\mathscr{L}$.

III. RESULTS AND DISCUSSION

A. Solid –liquid phase diagrams with a single eutectic point

In a first step, in order to test the validity and usefulness of this approach, we have studied several solid–liquid phase diagrams for systems that include a single eutectic point considering the idealizations and approximations stated in Sec. II. Thus, we have employed the methodology presented in Ref. 27 and have chosen some examples involving binary mixtures: Ag/Si, $KNO_3/LiNO_3$, and K_2SO_4/Li_2SO_4. The required data were obtained from the same source.[27]

For all cases, as a naive approach, we have computed grids in the (x_1, T) plane taking 100 points for x_1 in the [0, 1] range and 500 points for T in the $[0.25 \times \min(T_{f,1}, T_{f,2}), 1.25 \times \max(T_{f,1}, T_{f,2})]$ range such that the total number of grid points was 50 000. For the initialization set, we have taken another grid with 5 points for x_1 and 10 points for T in the same ranges as the former ones; then, we have sampled other 100 points using uncertainty sampling that gives a total number of 150 sampled points. In other words, in order to build a rather accurate phase diagram, we have used a **0.3% of points** compared to the naive all grid approach. The obtained phase diagrams are represented in Figs. 3–5.

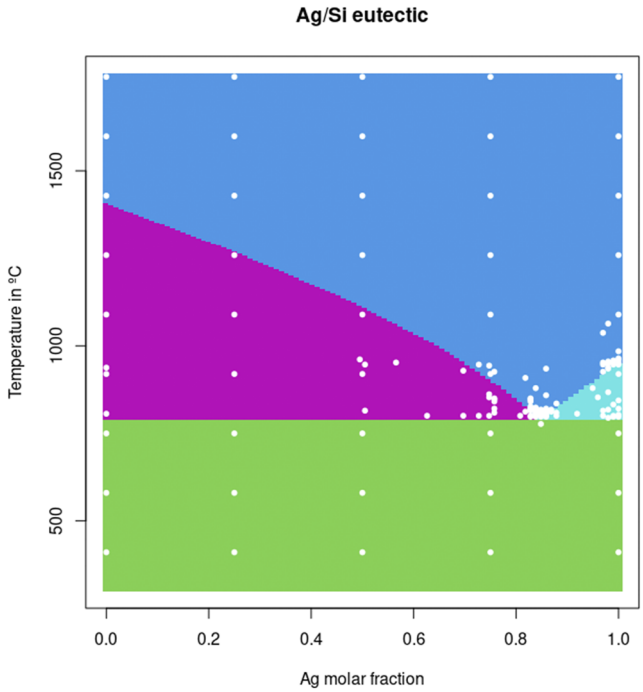

FIG. 3. Ag/Si solid–liquid phase diagram. Green is the heterogeneous solid phase $\alpha + \beta$, dark blue is the homogeneous ideal liquid mixture L, light blue is $\beta + L$, and purple is $\alpha + L$; the white dots represent the sampled points.

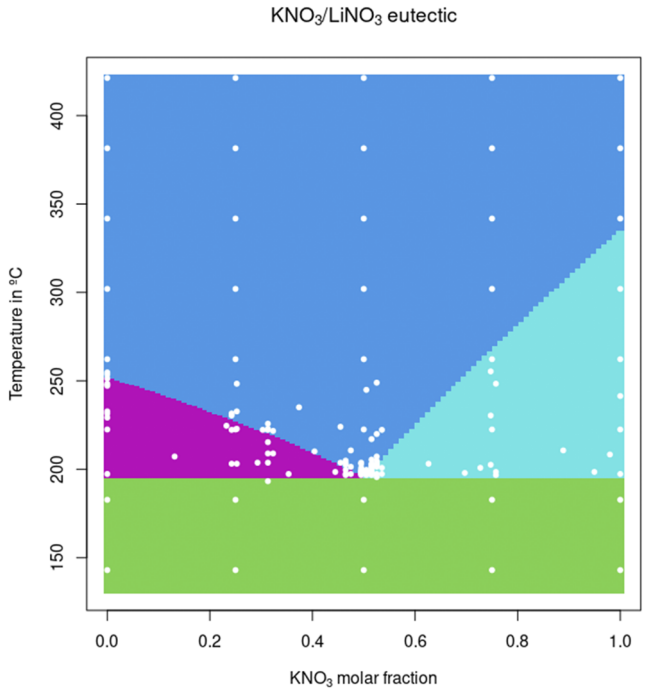

FIG. 4. $KNO_3/LiNO_3$ solid–liquid phase diagram. Green is the heterogeneous solid phase $\alpha + \beta$, dark blue is the homogeneous ideal liquid mixture L, light blue is $\beta + L$, and purple is $\alpha + L$; the white dots represent the sampled points.

As it has been observed in all three studied cases, the uncertainty sampling method samples more thoroughly the region where the larger number of phases coexist, i.e., near the eutectic point. Finding eutectic compositions is crucial for some pharmacological[45–47] and green chemistry[48,49] applications. Therefore, this method shall be interesting for discovering novel eutectic formulations.

B. Ground state spin multiplicity for Hooke atoms

For these systems, we have employed a grid in the $(\lambda, \omega_{xy}^2, \omega_z^2)$ space with 15 points for each dimension in the domain $[0.05, 2.55] \times [0.20, 0.30] \times [0.20, 0.30]$, therefore a total number of 3375 grid points. In this case, we have taken just 58 points in the grid to be computed, which is the 1.72% of all points.

We shall start our discussion by describing the obtained ground state spin multiplicity phase diagrams. In Fig. 6, we find magnetic phase diagrams for 4 (left hand side) and 6 (right hand side) electron systems. The vertical axis represents the electron–electron interaction screening parameter λ, while the horizontal one represents the ratio between the confinement along the x/y axis and the z axis $\varepsilon = \frac{\omega_{x,y}^2}{\omega_z^2}$. We shall proceed to analyze the obtained diagrams.

On the one hand, for 4 electron systems, one shall observe that singlets (nonmagnetic states labeled in red) appear in regions for which $\omega_{x,y}^2$ is large, ω_z^2 is small, and, for some $(\omega_{xy}^2, \omega_z^2)$ points close to the boundary, they are more abundant as λ increases. Under the anisotropic potential induced by the fact that $\omega_{x,y} > \omega_z$, the one body energies along the z axis are smaller than in the x, y plane and in

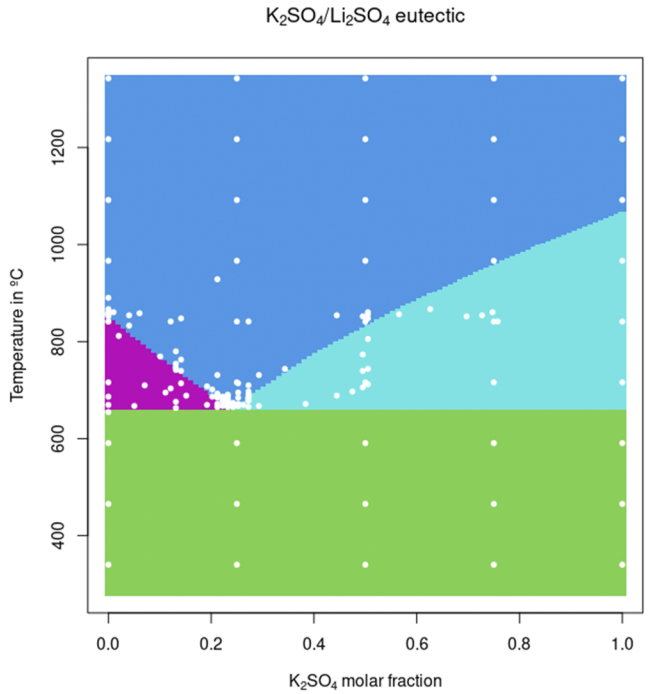

FIG. 5. K_2SO_4/Li_2SO_4 solid–liquid phase diagram. Green is the heterogeneous solid phase $\alpha + \beta$, dark blue is the homogeneous ideal liquid mixture L, light blue is $\beta + L$, and purple is $\alpha + L$; the white dots represent the sampled points.

virtue of the *Aufbau* principle, the former will be occupied first giving rise to singlet states. Given an anisotropic enough potential for which the ground state is still a triplet (as in the spherical case), we may still obtain a transition to the singlet state by increasing the screening potential λ as by doing so we are turning off the exchange interaction that does not compensate the high spin state and one body interactions will impose singlet spin states.

On the other hand, for 6 electron systems, it is obvious that some features of the phase diagram differ from the one obtained for the 4 electron system. First, the ground state spin multiplicity behaves in the opposite way as compared to the case for 4 electrons; in this case, singlet states appear for small values for $\omega_{x,y}^2$ curvatures and, at the same time, large values for ω_z^2. In addition, the behavior with respect to the electron interaction screening parameter λ is the same as in the previous case, the higher it is [for a couple of suitable curvatures $(\omega_{x,y}^2 < \omega_z^2)$], the most likely it is for that potential to give rise to a singlet spin multiplicity for the electronic ground state.

In both cases, for large values of the Yukawa-like screening parameter in the electron–electron interaction λ, quantum exchange interaction becomes weaker and both systems show ground state spin multiplicity transition in the close neighborhood of the vertical line $\varepsilon = 1$. In addition, in the low λ values regime where exchange interaction are not negligible, in both cases, we may observe that spin transition does not happen as soon as we distort the spherical symmetry of the potential; we observe there is an inertia to hold the high spin state (triplet $S = 1$ against singlet $S = 0$). This is a natural consequence of exchange interaction as it lowers the total energy of the

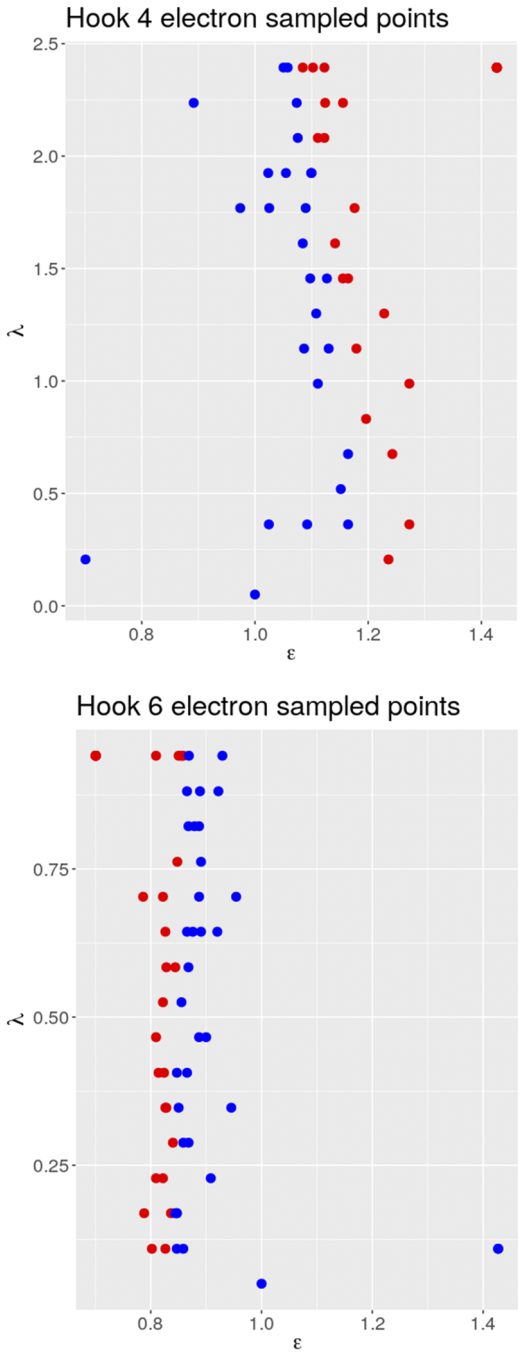

FIG. 6. Sampled points in the (ε, λ) plane for 4 (top) and 6 (bottom) electron systems. Red indicates singlet ground spin state, while blue indicates triplet ground spin state.

system via same spin particles; thus, the more particles with same spin, the more stable the system becomes.

Taking it up to the non-interacting electron system $(\lambda \to \infty)$, the Hamiltonian is only composed by one-particle operators for

which the orbital energy eigenvalues in atomic units are given by 8. In this case, depending on the asymmetry parameter ε, which is defined as the ratio between the axial confinement parameters $\omega_{x,y}^2$ and ω_z^2, the energy of the p orbitals will split as shown in Fig. 7. Taking into account the Pauli's exclusion principle, we may fit up to a couple of electrons in each orbital, and, according to Hund's rule, in case of degeneracy, they will occupy degenerate orbitals such that total spin is maximized. Keeping these two many-body quantum rules in mind, it is easy to see that the ground state spin multiplicity does not only depend on the confinement potential but also in the number of electrons,

$$\varepsilon_{n_x,n_y,n_z} = \omega_{x,y}(n_x + n_y + 1) + \omega_z\left(n_z + \frac{1}{2}\right). \tag{8}$$

When the confinement is sharper along the z direction than in the x and y directions ($\omega_z^2 > \omega_{x,y}^2$) (this is sphericity parameter $\varepsilon < 1$), the degeneracy of the p orbital splits into two main groups, the p_z orbital (higher in energy) and the plane composed by the still degenerate p_x and p_y (lower in energy). Therefore, when we distort the spherical symmetry in this way, nothing essential happens to 4 electron systems as far as ground state spin multiplicity is concerned. However, for six electron systems, as ε gets smaller, the energy of the p_z orbital goes higher and we obtain a point in which all p_x and p_y orbitals are doubly occupied while p_z orbital is empty. If we are to fit all 6 electrons in this potential, we shall see that the ground state spin multiplicity is the singlet; thus, magnetic properties are altered in the process of comprising the sphere.

On the contrary, instead of compressing the sphere if we stretch it along the z axis, we obtain a smaller curvature along this direction ($\omega_z^2 < \omega_{x,y}^2$) and the sphericity parameter becomes $\varepsilon > 1$. Once again, according to the one-body energy formula (8), this implies that the orbital energies, degenerating in the spherical case, split into two branches: the p_z orbital (this time lower in energy) and the plane composed by the degenerating p_x and p_y (higher in energy). Applying the occupancy rules, one shall see that, for 4 electron systems, the ground state multiplicity is the singlet (all paired electrons), while, for 6 electron systems, there is no change as long as ground state multiplicity is concerned. As opposed to the previous case, the magnetic properties for the 4 electron system alter along this transformation, while the 6 electron system is able to keep its magnetic properties.

C. Detecting covalent bonding in spherically confined He_2 systems

In a first step, we have built the covalency phase diagrams shown in Fig. 8 at HF/aug-cc-pVTZ level using as label the sign (red for negative and blue for positive) of the Laplacian of the density at the bond critical point $\Delta\rho(r_C)$ and the total energy density at the bond critical point $H(r_C)$. In total, we have computed 500 × 500 grid points in the $[0.10, 4.00] \times [1.00, 4.00]$ interval in the (ω^2, R) plane taking 104 points for initialization and 200 more for building the diagram, therefore, **using 0.1216% of all grid points**.

In these diagrams, we represent the confinement curvature ω^2 in the vertical axis and the interatomic distance R in the horizontal one. As it can be seen for $\Delta\rho(r_C)$, negative values are obtained in regions where the interatomic distance lays around 1.00 and 1.52 Å and confinement curvatures larger than 2.00. On the other hand, as far as $H(r_C)$ is concerned, we have found negative values for all interatomic distances shorter than 1.50 Å. Beyond this threshold, we may observe that, for large enough harmonic confinement curvature, the sign of the total energy density at the BCP switches from positive to negative; the curvature required to make this switch happen is larger as we increase the interatomic distance between the He atoms.

At this level of theory, for the range in which there are covalency indicators, our model comes in terms with previous studies based on all-atom calculations on similar systems. For example, for $He_2@C_{20}H_{20}$, the He–He distance is reported to be 1.265Å and positive value for $\Delta\rho(r_C)$[18] is compatible with our calculations for $\omega^2 < 2.00$. Also for interatomic distances smaller than 1.60Å, for $He_2@B_{12}N_{12}$ and $He_2@B_{16}N_{16}$ systems, 1.306 Å and 1.456 Å distances are reported.[24] For these two systems, positive $\Delta\rho(r_C)$ are reported, while total energy density is negative in the former (around 1×10^{-2} a.u.) and zero in the latter that comes along with our results. For larger distances, in a recent study, 1.520 Å and 1.546 Å distances are reported for $He_2@C_{36}$ and $He_2@C_{40}$, respectively;[20] the reported sign of $\Delta\rho(r_C)$ and $H(r_C)$ are positive (of the order of 5×10^{-3} a.u.). Finally, He atom couples confined in B_{40} cages have been reported

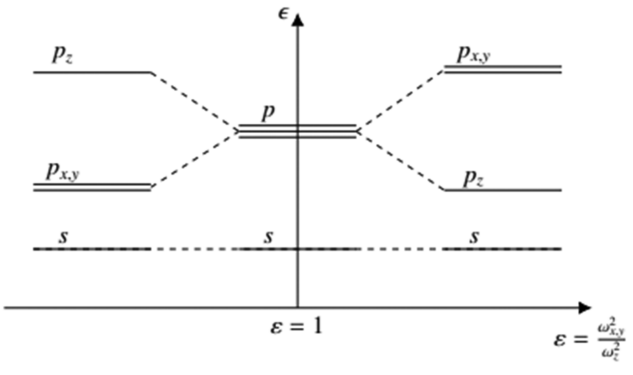

FIG. 7. Schematic atomic orbital energy splitting along asymmetry parameter ε.

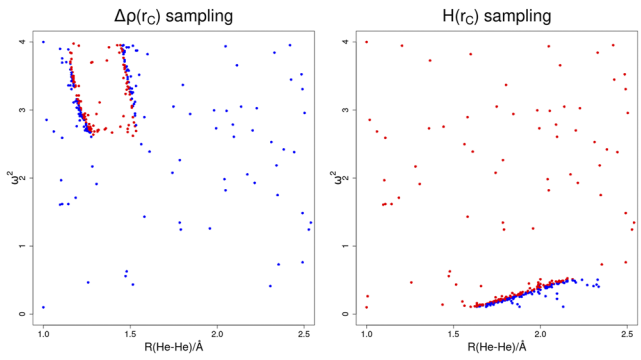

FIG. 8. Sign of total energy density for He_2 system computed at HF/aug-cc-pVTZ level: red is for negative and blue for positive.

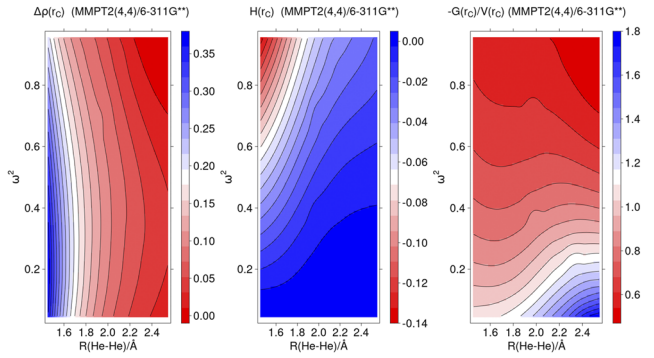

FIG. 9. Covalency indicators for MMPT2(4,8)/aug-cc-pVTZ calculated densities for spherical confinement. From left to right: $\Delta\rho(r_C)$, $H(r_C)$, and $-G(r_C)/V(r_C)$. Red shades indicate low values, white indicates medium, and blue shades indicate high values.

to be 1.672 – 1.640 $\AA$ apart,[25] and $\Delta\rho(r_C)$ and $H(r_C)$ both turned out to be positive, which is compatible with our results for small ω^2 values.

So far, results concerning the nature of bonding in helium dimer obtained by our harmonic confinement model seem to agree with all atom models for several cages. Bearing in mind that real cages may induce a rather small potential curvature ω^2 and the fact that switch in the sign of $H(r_C)$ happens somewhere in the interval $[1.40, 2.60]$ $\AA$ for internuclear distance, we shall focus our attention to studying this region and include electron-correlation effects in our calculations. Hence, we have performed CASSCF(4,8)/aug-cc-pVTZ calculations in this region. To do so, we have taken a 20×20 grid points in the (ω^2, R) plane with evenly separated points in the $(0.00, 1.00) \times (1.40, 2.6)$ domain. Using these points, we have computed level maps as in Fig. 9. In the three maps, we indicate the interatomic distance in the horizontal axis, the confinement curvature in the vertical axis, and we have computed level maps for three Bader descriptors (from left to right), $\Delta\rho(r_C)$, $H(r_C)$, and $-G(r_C)/V(r_C)$, which take values in different range according to each descriptor labeled in red for low values and in blue for higher ones. As we can see for $\Delta\rho(r_C)$, we have found positive values in this domain and it increases as the interatomic distance is shorter and it is not highly dependent on ω^2, especially for rather short interatomic distances. For $H(r_C)$, we have found negative values that get smaller (more negative) as we shorten the internuclear distance and increase the confinement curvature that is compatible with prior HF calculations as well as with all-atom calculations in which more negative values for this parameter appear when the size of the confining cage is smaller.

IV. CONCLUDING REMARKS

In this work, we have been able to produce computationally affordable magnetic and covalency phase diagrams for some few electron Hooke atoms and helium dimers confined in harmonic potentials, respectively, by sampling the potential feature space by means of semi-supervised learning technologies. Contrary to the conventional way of explicitly calculating every point in a grid, we have taken several initialization points and have propagated their information using classifiers based on Random Forest, which, compared to label propagation algorithm, has enabled us to study dense grids. By just calculating the points where information is maximal, we have been able to reduce the number of required computations (in our case, the required points for computation were below 2% of all grid points).

The validation of the method employing RFs was performed by computing analytical ideal solid–liquid phase diagrams involving a single eutectic point. As it has been shown, this method samples all interphases, especially the neighborhood of the eutectic point in which all four phases coexist.

From previous studies on 4 and 6 electron systems confined in spherical harmonic potentials with curvature $\omega^2 = 0.25$, it is known that the ground state spin multiplicity of these systems is a triplet ($S = 1$) and the first singlet ($S = 0$) lays around 1 eV above in energy. If we consider the electrons do not interact among them, the whole system behaves as a system composed by n bodies that still must fulfill basic many-body systems features as Pauli's exclusion principle, *Aufbau* filling principle, and Hund's rule. Under these assumptions, we state that ground state spin multiplicity will be imposed by the symmetry of the confining potential and, by altering the symmetry of it, we may obtain ground state multiplicities, either singlet or triplets, depending on the potential parameters, the strength of interelectronic interactions, and the number of electrons.

On the other hand, Bader analysis performed upon confined helium dimers by means of harmonic potential seems to be a rather good model that captures the main features of all atom approaches. As far as sign of density Laplacian and total energy density is concerned, all trends have been correctly described by this simple model that shows that effective confinement potentials can be employed to capture the essence of complex confined systems and rather simple models can be employed to describe them. This fact enables the usage of high theoretical level computations upon these systems.

As a first step for many other applications of chemical interest, we conclude these machine learning techniques may be useful for classification and exploration.

ACKNOWLEDGMENTS

This work has been carried out in the Theoretical Chemistry Group http://www.ehu.eus/chemistry/theory/category/1_group/ in the Faculty for Chemical Sciences of the University of the Basque Country and the Donostia International Physics Center (DIPC) in the frame of the project from the Basque Government GV IT1254-19 (October 07, 2019). The SGI/IZO-SGIker UPV/EHU is gratefully acknowledged for generous allocation of computational resources.

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

Author Contributions

Xabier Telleria-Allika: Conceptualization (equal); Formal analysis (equal); Investigation (equal); Methodology (equal); Writing-original-draft (equal); Writing-review-editing (equal). **Jose M. Mercero:** Supervision (equal); Writing-review-editing

(equal). **Xabier Lopez:** Data curation (equal); Methodology (equal); Resources (equal); Writing-review-editing (equal). **Jon M. Matxain:** Conceptualization (equal); Investigation (equal); Methodology (equal); Resources (equal); Supervision (equal); Writing-review-editing (equal).

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

[1] R. J. N. Baldock, L. B. Pártay, A. P. Bartók, M. C. Payne, and G. Csányi, "Determining pressure-temperature phase diagrams of materials," Phys. Rev. B **93**, 174108 (2016).

[2] Y.-C. Lo, S. E. Rensi, W. Torng, and R. B. Altman, "Machine learning in chemoinformatics and drug discovery," Drug Discovery Today **23**, 1538–1546 (2018).

[3] S. Gugler, J. P. Janet, and H. J. Kulik, "Enumeration of de novo inorganic complexes for chemical discovery and machine learning," Mol. Syst. Des. Eng. **5**, 139–152 (2020).

[4] F. Piroozmand, F. Mohammadipanah, and H. Sajedi, "Spectrum of deep learning algorithms in drug discovery," Chem. Biol. Drug Des. **96**, 886–901 (2020).

[5] J. P. Janet, L. Chan, and H. J. Kulik, "Accelerating chemical discovery with machine learning: Simulated evolution of spin crossover complexes with an artificial neural network," J. Phys. Chem. Lett. **9**, 1064–1071 (2018).

[6] T. F. G. G. Cova and A. A. C. C. Pais, "Deep learning for deep chemistry: Optimizing the prediction of chemical patterns," Front. Chem. **7**, 809 (2019).

[7] O. T. Unke and M. Meuwly, "Machine learning potential energy surfaces," arXiv:1909.08027 [physics.chem-ph] (2019).

[8] G. Schmitz, I. H. Godtliebsen, and O. Christiansen, "Machine learning for potential energy surfaces: An extensive database and assessment of methods," J. Chem. Phys. **150**, 244113 (2019).

[9] P. O. Dral, A. Owens, A. Dral, and G. Csányi, "Hierarchical machine learning of potential energy surfaces," J. Chem. Phys. **152**, 204110 (2020).

[10] X. Xie, K. A. Persson, and D. W. Small, "Incorporating electronic information into machine learning potential energy surfaces via approaching the ground-state electronic energy as a function of atom-based electronic populations," J. Chem. Theory Comput. **16**, 4256–4270 (2020).

[11] O. T. Unke, D. Koner, S. Patra, S. Käser, and M. Meuwly, "High-dimensional potential energy surfaces for molecular simulations: From empiricism to machine learning," Mach. Learn. **1**, 013001 (2020).

[12] C. W. Coley, R. Barzilay, T. S. Jaakkola, W. H. Green, and K. F. Jensen, "Prediction of organic reaction outcomes using machine learning," ACS Cent. Sci. **3**, 434–443 (2017).

[13] H. Gao, T. J. Struble, C. W. Coley, Y. Wang, W. H. Green, and K. F. Jensen, "Using machine learning to predict suitable conditions for organic reactions," ACS Cent. Sci. **4**, 1465–1476 (2018).

[14] T. D. Rhone, R. Hoyt, C. R. O'Connor, M. M. Montemore, C. S. S. R. Kumar, C. M. Friend, and E. Kaxiras, "Predicting outcomes of catalytic reactions using machine learning," arXiv:1908.10953 [physics.chem-ph] (2019).

[15] K. Terayama *et al.*, "Efficient construction method for phase diagrams using uncertainty sampling," Phys. Rev. Mater. **3**, 033802 (2019).

[16] C. Dai and S. C. Glotzer, "Efficient phase diagram sampling by active learning," J. Phys. Chem. B **124**, 1275–1284 (2020).

[17] Z. G. X. Zu and J. Lafferty, "Learning from labeled and unlabeled data with label propagation," in *Proceedings of the 20th International Conference on Machine Learning* (School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, 2003), pp. 912–919.

[18] E. Cerpa, A. Krapp, R. Flores-Moreno, K. J. Donald, and G. Merino, "Influence of endohedral confinement on the electronic interaction between He atoms: A $He_2@C_{20}H_{20}$ case study," Chem. - Eur. J. **15**, 1985–1990 (2009).

[19] T. Y. Nikolaienko and E. S. Kryachko, "Formation of dimers of light noble atoms under encapsulation within fullerene's voids," Nanoscale Res. Lett. **10**, 185 (2015).

[20] D. Paul, H. Dua, and U. Sarkar, "Confinement effects of a noble gas dimer inside a fullerene cage: Can it be used as an acceptor in a DSSC?," Front. Chem. **8**, 621 (2020).

[21] A. Krapp and G. Frenking, "Is this a chemical bond? A theoretical study of $Ng_2@C_{60}$ (Ng=He, Ne, Ar, Kr, Xe)," Chem. - Eur. J. **13**, 8256–8270 (2007).

[22] S. Gómez and A. Restrepo, "Noble gas dimers confined inside C_{70}," Phys. Chem. Chem. Phys. **21**, 15815–15822 (2019).

[23] S. Jalife, J. Arcudia, S. Pan, and G. Merino, "Noble gas endohedral fullerenes," Chem. Sci. **11**, 6642–6652 (2020).

[24] M. Khatua, S. Pan, and P. K. Chattaraj, "Confinement induced binding of noble gas atoms," J. Chem. Phys. **140**, 164306 (2014).

[25] S. Pan, M. Ghara, S. Kar, X. Zarate, G. Merino, and P. K. Chattaraj, "Noble gas encapsulated B_{40} cage," Phys. Chem. Chem. Phys. **20**, 1953–1963 (2018).

[26] D. Chakraborty and P. K. Chattaraj, "Confinement induced binding in noble gas atoms within a BN-doped carbon nanotube," Chem. Phys. Lett. **621**, 29–34 (2015).

[27] L. Brunet, J. Caillard, and P. André, "Thermodynamic calculation of n-component eutectic mixtures," Int. J. Mod. Phys. C **15**, 675–687 (2004).

[28] M. Taut, "Two electrons in an external oscillator potential: Particular analytic solutions of a Coulomb correlation problem," Phys. Rev. A **48**, 3561–3566 (1993).

[29] M.-A. Martínez-Sánchez *et al.*, "Shannon entropy for the hydrogen atom confined by four different potentials," Quantum Rep. **1**, 208–218 (2019).

[30] H. Gholizadehkalkhoran *et al.*, "Eigenstates and dynamics of Hooke's atom: Exact results and path integral simulations," J. Math. Phys. **59**, 052104 (2018).

[31] J. Katriel *et al.*, "Hund's rule in open-shell states of two-electron systems: From free through confined and screened atoms, to quantum dots," Nanosystems **10**, 31–41 (2019).

[32] M. Slamet *et al.*, "Kinetic effects in 2D and 3D quantum dots: Comparison between high and low electron correlation regimes," Comput. Theor. Chem. **1138**, 140–157 (2018).

[33] A. Odriazola *et al.*, "Universal scaling relations for the energies of many-electron Hooke atoms," Phys. Rev. A **95**, 042511 (2017).

[34] C. A. Downing, "Two-electron atom with a screened interaction," Phys. Rev. A **95**, 022105 (2017).

[35] J. M. Ugalde and C. Sarasola, "Evaluation of screened nuclear attraction and electron repulsion molecular integrals over Gaussian basis functions," Int. J. Quantum Chem. **62**, 273–278 (1997).

[36] X. Telleria-Allika, J. M. Ugalde, X. Lopez, E. Matito, E. Ramos-Cordoba, and M. Rodriguez-Mayorga, "Aufbau principle and singlet-triplet gap in spherical Hooke atoms," Int. J. Quant. Chem. (submitted) (2022).

[37] K. Varga, P. Navratil, J. Usukura, and Y. Suzuki, "Stochastic variational approach to few-electron artificial atoms," Phys. Rev. B **63**, 205308 (2001).

[38] K. Strasburger, "The order of three lowest-energy states of the six-electron harmonium at small force constant," J. Chem. Phys. **144**, 234304 (2016).

[39] J. Cioslowski, K. Strasburger, and E. Matito, "Benchmark calculations on the lowest-energy singlet, triplet, and quintet states of the four-electron harmonium atom," J. Chem. Phys. **141**, 044128 (2014).

[40] A. Robles-Navarro, M. Rodriguez-Bautista, P. Fuentealba, and C. Cárdenas, "The change in the nature of bonding in the Li_2 dimer under confinement," Int. J. Quantum Chem. **121**, e26644 (2021).

[41] B. Settles, *Active Learning* (Morgan and Claypool, 2012).

[42] B. Settles, "Active learning literature survey," Computer Sciences Technical Report No. 1648, University of Wisconsin–Madison, 2009.

[43] D. D. Lewis and W. A. Gale, "A sequential algorithm for training text classifiers," in *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, 49 (AT&T Bell Laboratories, Murray Hill, NJ, 1994), pp. 3–12.

[44] C. E. Shannon, "A mathematical theory of communication," Bell Syst. Tech. J. **27**, 623–656 (1948).

[45] I. Balakrishnan, N. Jawahar, S. Venkatachalam, and D. Datta, "A brief review on eutectic mixture and its role in pharmaceutical field," Int. J. Res. Pharm. Sci. **11**, 3017–3023 (2020).

[46]A. P. Abbott, E. I. Ahmed, K. Prasad, I. B. Qader, and K. S. Ryder, "Liquid pharmaceuticals formulation by eutectic formation," Fluid Phase Equilib. **448**, 2–8 (2017), part of special issue on Deep Eutectic Solvents.
[47]C. Huang, X. Chen, C. Wei, H. Wang, and H. Gao, "Deep eutectic solvents as active pharmaceutical ingredient delivery systems in the treatment of metabolic related diseases," Front. Pharmacol. **12**, 794939 (2021).
[48]Y. Chen and T. Mu, "Revisiting greenness of ionic liquids and deep eutectic solvents," Green Chem. Eng. **2**, 174–186 (2021).
[49]L. P. Silva, M. A. R. Martins, J. H. F. Conceição, S. P. Pinho, and J. A. P. Coutinho, "Eutectic mixtures based on polyalcohols as sustainable solvents: Screening and characterization," ACS Sustainable Chem. Eng. **8**, 15317–15326 (2020).