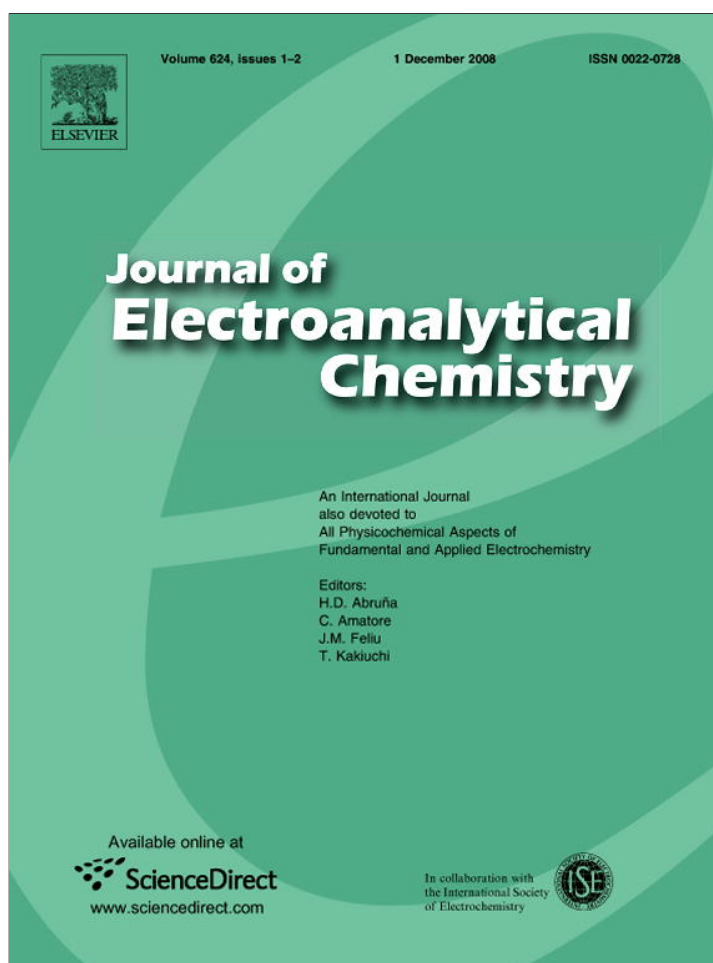


Provided for non-commercial research and education use.
Not for reproduction, distribution or commercial use.



This article appeared in a journal published by Elsevier. The attached copy is furnished to the author for internal non-commercial research and education use, including for instruction at the authors institution and sharing with colleagues.

Other uses, including reproduction and distribution, or selling or licensing copies, or posting to personal, institutional or third party websites are prohibited.

In most cases authors are permitted to post their version of the article (e.g. in Word or Tex form) to their personal website or institutional repository. Authors requiring further information regarding Elsevier's archiving and manuscript policies are encouraged to visit:

<http://www.elsevier.com/copyright>



Contents lists available at ScienceDirect

Journal of Electroanalytical Chemistry

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

Mapping Potential Energy Surfaces by Neural Networks: The ethanol/Au(1 1 1) interface

Diogo A.R.S. Latino^{a,b}, Rui P.S. Fartaria^a, Filomena F.M. Freitas^a, João Aires-de-Sousa^b,
Fernando M.S. Silva Fernandes^{a,*}^a Centre of Molecular Sciences and Materials, Department of Chemistry and Biochemistry, Faculty of Sciences, University of Lisboa, Campo Grande, 1749-016 Lisboa, Portugal^b CQFB and REQUIMTE, Department of Chemistry, Faculty of Sciences and Technology, New University of Lisboa, 2829-516 Caparica, Portugal

ARTICLE INFO

Article history:

Received 9 June 2008

Received in revised form 29 July 2008

Accepted 31 July 2008

Available online 11 September 2008

Keywords:

Potential Energy Surfaces

Ensembles of Feed-Forward Neural

Networks

Ethanol adsorption on Au(1 1 1)

ABSTRACT

Potential Energy Surfaces (PES) for the ethanol/Au(1 1 1) interface are mapped by Neural Networks (NNs). Interaction energies, calculated from Density Functional Theory (DFT), for the adsorption of the ethanol on Au(1 1 1) surfaces are used to train Ensembles of Feed-Forward Neural Networks (EnFFNNs). The distance of the ethanol molecule to the surface, two angles describing the molecular orientation relatively to the surface, and three binary descriptors encoding the gold adsorption sites, are the input to the NNs. The training sets contain energy values at different distances, for seven molecular orientations and three adsorption sites. The models are assessed by: (a) internal cross validation; (b) Leave-One-Out procedure (LOO); and (c) external test sets corresponding to orientations not used in the training procedure. The results are compared with the ones obtained from an analytical force field recently proposed by some of us to match the DFT data. It is shown that NNs can be trained to map PES with a similar or better accuracy than analytical representations. This is a relevant point, particularly in simulations by Monte Carlo (MC) or Molecular Dynamics (MD), which require an extensive screening of the interaction sites at the interface, turning the development of analytical functions a non-trivial task as the complexity of the systems increases.

© 2008 Elsevier B.V. All rights reserved.

1. Introduction

The structure and dynamics of electrode/solution interfaces are of great importance in the domain of electrochemistry. The modification of metallic surfaces properties by the adsorption of molecules allows, for example, photovoltaic, biosensing, and corrosion protection developments.

The adsorption and spontaneous organization of organic molecules on metallic surfaces giving rise to films of organized and stable monolayers is known as Self-Assembled Monolayers (SAMs). SAMs can be produced using different types of molecules and substrates. Typically, alkane chains (with 10 or more methylene units) and a thiol (SH) head group are used with Au surfaces due to the well-known chemical affinity between sulfur and gold. These species have the advantage and singular characteristic of creating dense monolayers when the thiol molecules adsorb onto gold with the tail chains pointing outwards the surface. Moreover, it is possible to functionalize the tail chain, after the formation of the SAMs, by the chemical insertion of specific functional groups or molecules.

To investigate the mechanisms involved in the adsorption and self-assembly of solvated organic molecules on metallic electrodes by Monte Carlo (MC) or Molecular Dynamics (MD) simulations, the determination of the Potential Energy Surfaces (PES) of the systems is crucial. They should describe the interactions between the molecular species present in the liquid phase as well the interactions between those species and the electrodes. Our current interest is the study of the adsorption of alkylthiols, solvated by ethanol, on gold electrodes and the understanding of the physi- and chemisorption mechanisms. To this end, our group has recently proposed an analytical force field, based on Density Functional Theory (DFT) calculations, for the interaction of ethanol with Au(1 1 1) surfaces. A preliminary test of the force field has also been carried out by MC simulations [1].

A function that matches DFT data provides, on one hand, a topographical visualization of the surface features, which may not be evident from a coarse-grained quantum mechanical study. On the other hand, it is a suitable input for simulation work. A good representation of PES should smoothly connect the asymptotic as well as the most interactive regions of the configuration space. It should accurately represent the true potential energy in the regions for which experimental or theoretical results are available and predict the interaction energies for the regions where such

* Corresponding author. Tel.: +351 21 7500133; fax: +351 21 7500088.

E-mail address: fsilva@fc.ul.pt (F.M.S.S. Fernandes).

data is not available. Fitting analytical functions to the energies of a set of suitable configurations of the system is one of the standard approaches to obtain PES. The London–Eyring–Polanyi–Sato (LEPS) functions, many-body expansions, splines and semiempirical potentials with adjustable parameters to reproduce experimental and theoretical results, are commonly used [2,3].

It appears, however, that as the complexity of the systems increases, the development of accurate analytical functions becomes a non-trivial task. In the last years, Neural Networks (NNs) turned out as an alternative way for mapping PES from *ab initio*/DFT energy data sets. [4–15] In such approximation, there are no *a priori* guesses of analytical functions and the results come out in tabular form. Moreover, once the networks are well trained, they are able to produce, as output, any required number of energy points for numerical interpolations with similar or better accuracy than other representation methods.

Single Feed-Forward Neural Networks (FFNNs) have been applied, for example, to obtain PES for water dimer [4], HCl^+ [5], OH^+ [6] and H^+ [7].

Some authors have compared structural, dynamical and thermal properties, obtained by MC and MD simulations, from NNs – PES and analytical functions. Gassner et al. [8] applied FFNNs to reproduce the three body interaction energy of the system $\text{H}_2\text{O}-\text{Al}^{3+}-\text{H}_2\text{O}$ and assessed the MC radial distribution functions, using an analytical function and the NN-PES. A similar study was done

by Cho et al. [9] where a polarizable force field for water was developed using NNs and tested in MC simulations.

A more systematic work was performed by Witkoskie and Doren [10] who studied the NNs accuracy in terms of some parameters such as the optimal number of neurons and data needed in the training. Likewise, our group [11] trained ensembles of Feed-Forward Neural Networks (EnsFFNNs) and Associative Neural Networks (ASNNs) for mapping PES also represented by well-known analytical potential functions, and then assessed the accuracy of the method by comparison of the simulation results from NNs and analytical PES. Training sets with different number of points, from 15 differently parametrized Lennard–Jones (LJ) potentials, were used and argon was taken to test the models. MD simulations were performed using the tabular potential energies, predicted by the NNs, to work out thermal, structural and dynamical properties which compare well with the values obtained from the LJ analytical function.

NNs have also been used to obtain PES for the study of adsorption on metallic surfaces. The study of reactions in surfaces was performed by Lorenz et al. [12] to build continuous PES for H_2 interacting with a (2×2) Pd(100) surface covered by potassium. More recently, the same authors [13] applied NNs to fit six dimensional PES for H_2 dissociation on the clean and sulfur covered Pd(100) surfaces. The models describe reaction rates for the dissociative adsorption and show that a description of dissociation reac-

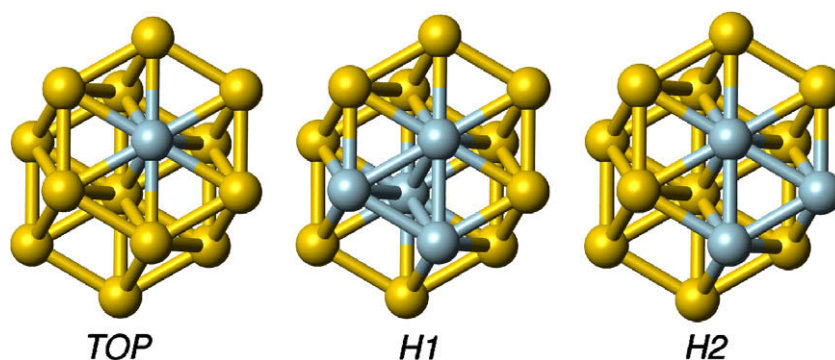


Fig. 1. Surface sites, top, hollow1 and hollow2, chosen to set up the ethanol–Au surface interaction.

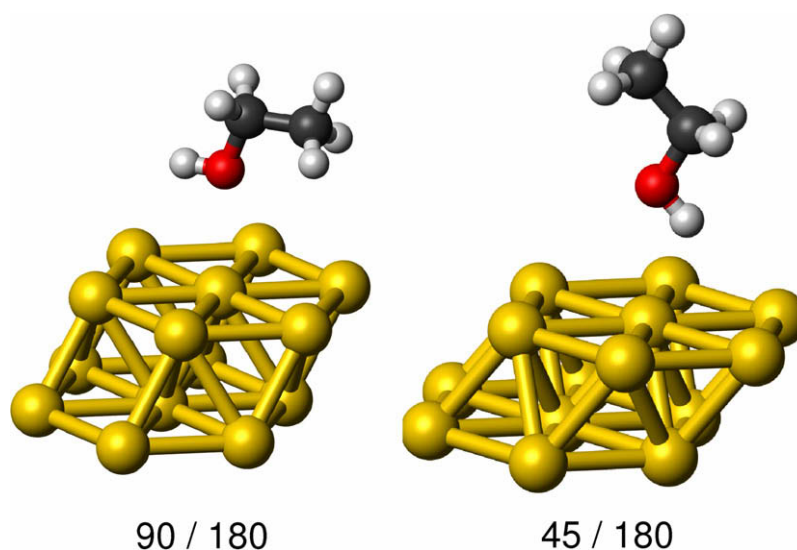


Fig. 2. Two of the orientations of the ethanol molecule relative to the Au(111) surface used in the training set (notation: α/β , see text).

tions with NNs is orders of magnitude smaller than that of the one for “on the fly” *ab initio* dynamics.

The main objective of the present work is to assess an alternative to our analytical force field [1], in order to map multidimensional PES for the interaction of ethanol and Au(111) surfaces regarding the simulation of the adsorption and self-assembly of alkythiols solvated by ethanol.

The machine learning methods used to estimate the full energy surface have been EnsFFNNs [16,17]. They show a greater generalization ability over single FFNNs in problems of modelling and fitting. EnsFFNNs, a supervised learning technique, are an extension of single FFNNs and a memory-less method (after the

training, all information about the input patterns are stored in the NN weights without the explicit storage of the data in the system).

The distance of the ethanol molecule to the surface, two angles describing the molecular orientation relatively to the surface, and three binary descriptors encoding the gold adsorption sites, are the input to the NNs. The training sets contain energy values at different distances, for seven molecular orientations and three adsorption sites. The models are assessed by: (a) internal cross validation; (b) Leave-One-Out procedure (LOO); and (c) external test sets corresponding to orientations not used in the training procedure.

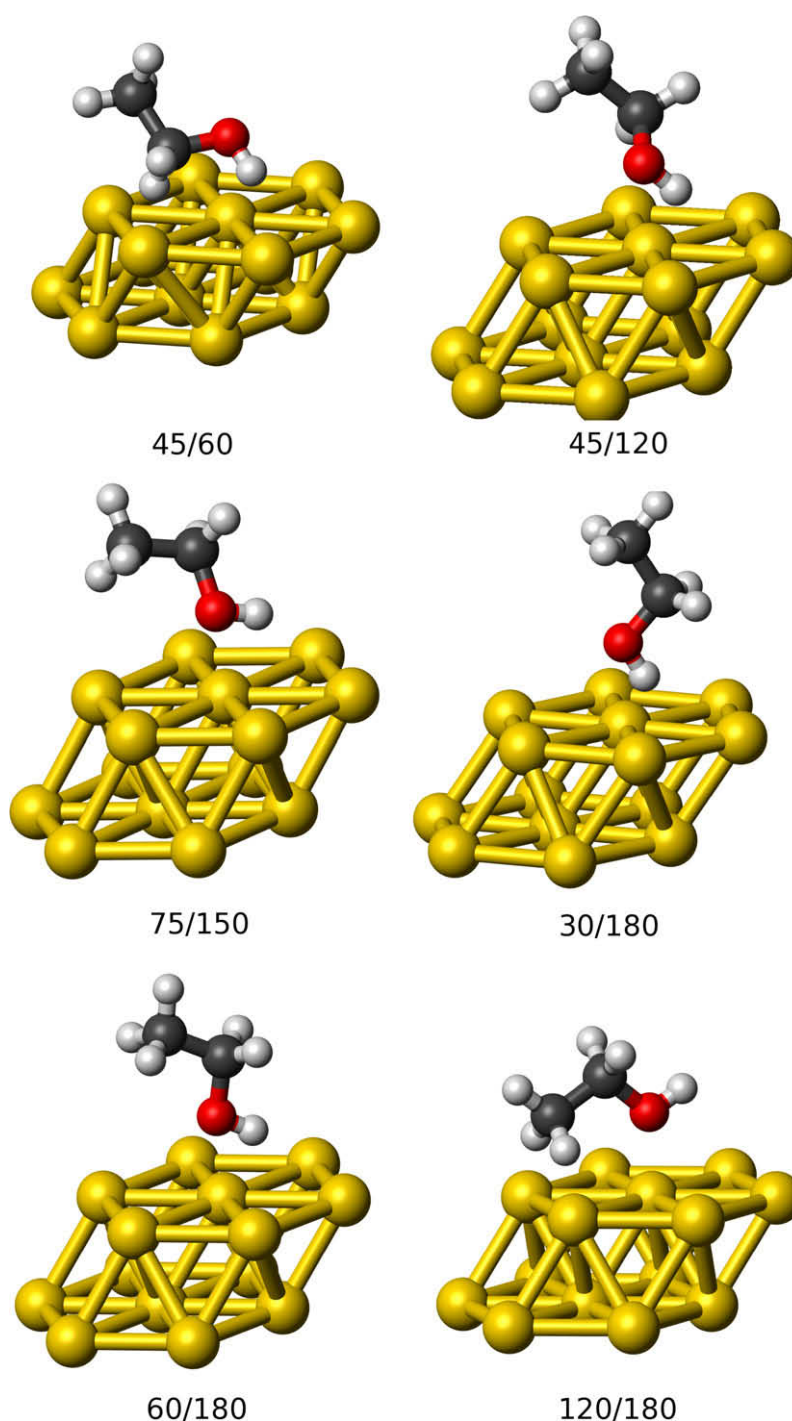


Fig. 3. Selected orientations of the ethanol molecule relative to the Au(111) surface used in the test set (notation: α/β , see text).

The next section contains the methodology and computational details. Section 3 discusses the NNs results and their comparison with the ones from the analytical force field. Section 4 presents the concluding remarks.

2. Methodology and computational details

The training has been performed with the DFT data of Fartaria et al. [1] used to set up the analytical force field. Additional DFT en-

ergy points have been calculated, at six orientations, to test the accuracy of the model in regions of the PES not considered by the training set.

EnsFFNNs is the machine learning method used to set up the relationship between the input (two orientation angles, the distance between the ethanol oxygen atom and the plane of the first layer of the Au(111) surface, and three binary descriptors to encode the adsorption sites) and the output (potential energy). The models are tested with different internal data sets (internal cross validation and LOO procedure) and external data sets.

Table 1
Energy minima from DFT, analytical function and EnsFFNNs training set

Ethanol orientation		Distance	Potential energy – U/kJ mol ⁻¹		
α /degrees	β /degrees	O-surface r/Å	DFT	Anal. function	EnsFFNNs
<i>Top site</i>					
0	180	3.54	7.76	8.7 (0.94)	7.96 (0.20)
135	180	4.07	5.87	6.7 (0.83)	6.15 (0.28)
180	180	6.0	1.37	1.2 (0.17)	1.78 (0.41)
45	180	3.3	9.04	15.0 (5.96)	9.13 (0.09)
90	0	5.5	2.32	1.5 (0.82)	2.38 (0.06)
90	180	2.67	19.03	19.25 (0.22)	18.17 (0.86)
90	90	3.54	11.77	10.4 (1.37)	11.95 (0.18)
<i>H1 site</i>					
0	180	3.3	13.14	10.7 (2.44)	12.93 (0.21)
135	180	4.29	4.34	5.5 (1.16)	3.96 (0.34)
180	180	6.04	1.35	0.9 (0.45)	1.51 (0.16)
45	180	2.87	16.42	13.3 (3.12)	14.61 (1.81)
90	0	5.57	2.35	1.4 (0.95)	2.46 (0.11)
90	180	2.87	12.8	15.5 (2.7)	13.91 (1.11)
90	90	3.87	5.99	7.1 (1.11)	6.03 (0.04)
<i>H2 site</i>					
0	180	3.3	10.46	9.8 (0.66)	10.32 (0.14)
135	180	4.2	5.08	5.3 (0.22)	4.96 (0.12)
180	180	6.04	1.48	1.0 (0.48)	1.42 (0.06)
45	180	3.08	11.82	12.7 (0.88)	11.39 (0.43)
90	0	5.37	2.33	1.4 (0.93)	2.37 (0.04)
90	180	2.67	15.07	14.7 (0.37)	14.86 (0.21)
90	90	3.87	5.75	7.1 (1.35)	6.03 (0.28)

In parenthesis is the absolute error.

Table 2
Energy minima from DFT, analytical function and EnsFFNNs test set

Ethanol orientation		Distance	Potential energy – U/kJ mol ⁻¹		
α /degrees	β /degrees	O-surface r/Å	DFT	Anal. function	EnsFFNNs
<i>Top site</i>					
45	60	3.79	7.0	8.78 (1.78)	-2.70 (9.70)
45	120	3.37	9.81	14.14 (4.33)	5.65 (4.16)
75	150	2.81	15.91	18.85 (2.94)	17.09 (1.18)
30	180	3.37	7.75	13.36 (5.61)	8.10 (0.35)
60	180	3.17	11.52	16.63 (5.11)	11.93 (0.41)
120	180	3.17	11.84	12.13 (0.29)	11.49 (0.35)
<i>H1 site</i>					
45	60	3.79	6.69	7.20 (0.51)	3.07 (3.62)
45	120	3.17	12.5	13.03 (0.53)	7.48 (5.02)
75	150	2.81	12.58	15.46 (2.88)	12.82 (0.32)
30	180	3.17	15.59	12.76 (2.83)	14.86 (0.73)
60	180	2.81	16.71	14.34 (2.37)	14.68 (2.03)
120	180	3.58	6.64	10.43 (3.79)	6.43 (0.21)
<i>H2 site</i>					
45	60	4.02	4.44	6.55 (2.11)	2.11 (2.33)
45	120	2.99	10.61	12.09 (1.48)	3.96 (6.65)
75	150	2.81	13.78	14.50 (0.72)	13.24 (0.54)
30	180	3.17	10.73	12.19 (1.46)	10.54 (0.19)
60	180	2.81	13.09	13.37 (0.28)	12.59 (0.50)
120	180	3.37	8.55	10.61 (2.06)	7.88 (0.67)

In parenthesis is the absolute error.

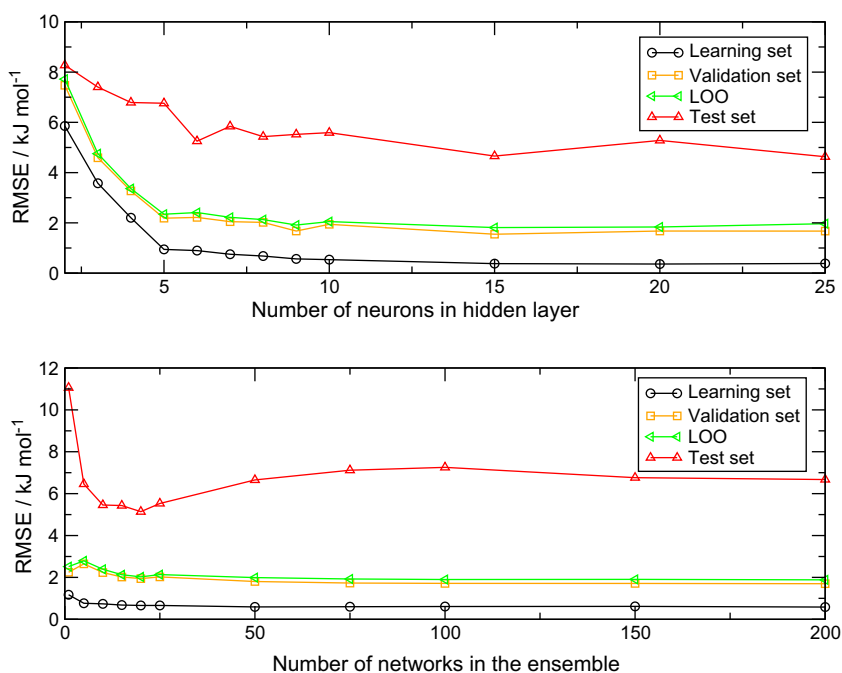


Fig. 4. Root Mean Square Error of the different data sets for networks with different number of neurons in the hidden layer (first graphic) and ensembles of different sizes (second graphic).

The results are compared with the values from DFT calculations and the ones from the analytical function developed. The full NNs-PES for each site in 3D representation is presented and analysed. A model is also trained with a different training and test sets, and with all the available data being compared with the one obtained from the initial training set. The interpolation ability of the NN is analysed.

2.1. DFT calculations and analytical force field

The theory level chosen to calculate the interaction energy of ethanol/Au(111) was the hybrid B3LYP method [18,19] with the LanL1MB basis set [20] applied to the gold atoms and the 6-31G basis set [21] for the H, C and O atoms. The calculations have been performed by the Gaussian 98 package [22].

A cluster of 14 Au atoms, to represent the gold surface, and one ethanol molecule, in the optimized gas-phase geometry, have been used to model the ethanol/Au(111) PES. The size of the gold cluster

was chosen with a compromise between the consistency of the interaction energy and the computation time. Fig. 1 shows the three adsorption sites, top (*Top*), hollow 1 (*H1*) and hollow 2 (*H2*), that have been selected to study the ethanol/Au(111) interaction.

In the *Top* site, the oxygen atom of ethanol approaches the surface directly over a gold atom of the first layer; the *H1* site corresponds to a hexagonal closed packed (hcp) site and the approach of the ethanol is made in the direction of the centre of a triangle formed between three gold atoms of the first layer with a gold atom of the second layer at the centre; and the *H2* site corresponds to a face centred cubic (fcc) site and the ethanol approach is made in the direction of the centre of a triangle formed between three gold atoms of the first layer.

The ethanol/Au(111)₁₄ cluster interaction energy, as a function of the distance and orientations of ethanol molecule to the Au(111) surface, is calculated by

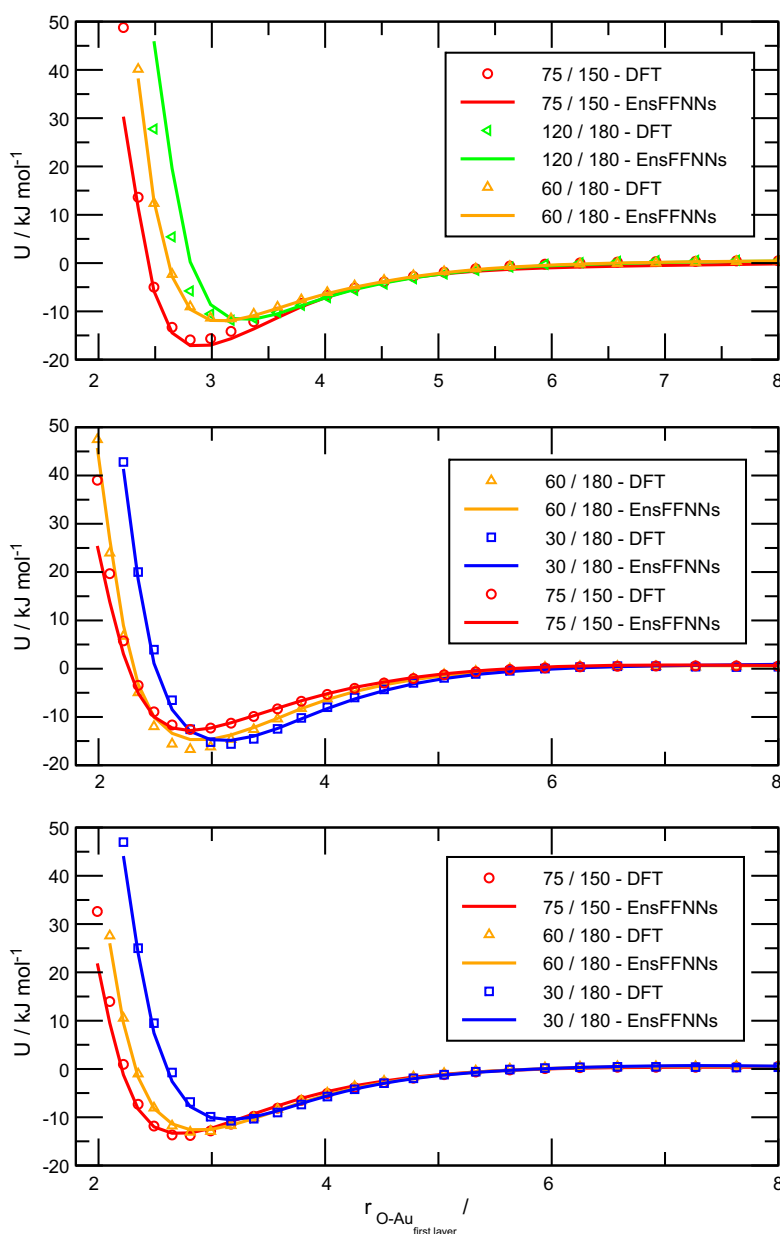


Fig. 5. Ethanol–Au(111)₁₄ potential energy curves by DFT and EnsFFNNs for the *Top*, *H1* and *H2* sites (from top to bottom, respectively) from the test set.

$$U_{\text{ethanol-Au}(111)_{14}}(r, \alpha, \beta) = U_{\text{Au}(111)+\text{ethanol}}(r, \alpha, \beta) - U_{\text{Au}(111)_{14}} - U_{\text{ethanol}} \quad (1)$$

where $U_{\text{Au}(111)+\text{ethanol}}$ is the energy of the system composed by the ethanol molecule and the cluster; $U_{\text{Au}(111)}$ and U_{ethanol} are the energies of the isolated cluster and ethanol molecule; r is the distance from the ethanol oxygen atom to the plane of the first layer of the Au(111) surface; α is the angle between the O–H bond and the normal to the surface and β is the angle between the plane H–O–C and the plane H–O–normal to the surface (both angles in degrees). The orientations for the ethanol molecule have been selected to span a wide range. Fig. 2 shows the two orientations with the highest binding energies used in the training, and Fig. 3 displays the molecular orientations used to test the models. The full list of orientations used in the training set are presented in Table 1 and the respective snapshots in Ref. [1].

For each orientation, several values of r , along the interval 0–10 Å, have been chosen and the $U_{\text{ethanol-Au}(111)_{14}}$ evaluated. Full details of the DFT calculations are described elsewhere [1].

The following analytical function was fitted to the above mentioned DFT results:

$$U_{\text{EtOH-Au}} = U_{\text{H}}(r_{\text{H-Au}}) + \left(1 + B_0 \cos(\theta/\text{rad})^{20} / r_{\text{O-Au}}^3\right) U_{\text{O}}(r_{\text{O-Au}}) + U_{\text{CH}_2}(r_{\text{CH}_2-\text{Au}}) + U_{\text{CH}_3}(r_{\text{CH}_3-\text{Au}}) + V(r_{\text{O-Au}}, \theta, \phi) \quad (2)$$

where

$$U_i(r_{i-\text{Au}}) = A_{0,i} \exp[A_{1,i}(r_{i-\text{Au}} + A_{2,i})] - A_{3,i} \exp[A_{4,i}(r_{i-\text{Au}} + A_{5,i})] \quad (3)$$

is the site–site interaction energy, and

$$V(r_{\text{O-Au}}, \theta, \phi) = C_0 \sin(\theta/\text{rad})^6 \exp[(r \sin(\theta/\text{rad}) - C_1)^2 / C_2] (C_3 - C_4 \cos(3(\phi/\text{rad} - C_5)) + C_6 \cos(6(\phi/\text{rad} - C_5))) \quad (4)$$

is the contribution due to the surface symmetry around a *Top* site. It is noteworthy that the cosine term with exponent 20, in Eq. 2, can be expressed by other forms (e.g. $\exp(-10x^2)$).

The function expresses the sum of the interactions between each gold atom and the sites of the ethanol molecule further modulated by two angular contributions, one related to the *Top* site surface symmetry and the other to the O–Au relative direction. The ethanol molecule was described by a united atom model with the H, O, CH₂ and CH₃ as the interaction sites. The variables used are: the distances, r_i , from each Au atom to each of the ethanol sites; the angle, θ , between the $r_{\text{O-Au}}$ vector and the normal to the surface; and the angle, ϕ , between the projection of the $r_{\text{O-Au}}$ vector on the surface plane and a reference surface vector beginning in a *Top* site and directed to a *H1* site. The fit to the DFT results was performed by means of genetic algorithms, using an Au(111) double layered electrode with 74 Au atoms in order to minimize border effects from the surface.

The overall fitting quality is good. We shall return to it in Section 3. The function parameters and the details of the fitting procedure can be seen elsewhere [1].

2.2. Feed-Forward Neural Networks

FFNNs [23] were implemented with six input neurons, one hidden layer of neurons, and one output neuron. In the input layer and in the hidden layer, an extra neuron (called bias) with the value of one was also added. A NN converts the input data X , $X = (x_1, x_2, \dots, x_i, \dots, x_m)$ into the output data Y , $Y = (y_1, y_2, \dots, y_i, \dots, y_n)$. Each neuron of the first layer receives all the input signals of an object, and the signals, after processing by the neuron, are sent to all neurons of the next layer. Each neuron of one layer is connected to all the neurons of the next layer. These connections are associated

with weights that represent the strength of the connection. All neurons j in the network perform three basic operations:

- (1) Obtain input signals from m neurons.
- (2) Convert these signals to a *Net* input signal using the expression:

$$\text{Net}_j = \sum_{i=1}^m w_{ji} x_i \quad (5)$$

where w_{ji} is the strength of the connection between neuron i and neuron j , and x_i is the input signal from neuron i .

- (3) Transform the *Net* signal into an output signal:

$$\text{Out} = f(\text{Net}_j) = \frac{1}{1 + \exp(-\gamma \text{Net}_j + v)} \quad (6)$$

The function f is called the transfer function which, in the present study, is a sigmoidal function.

The networks were trained using the ASNN program of Igor Tetko [24,25] taking as input the distance between the ethanol oxygen atom and the plane of the first layer of the Au(111) surface (r), two angles to describe the orientation of the ethanol to the surface (α and β) and three binary descriptors to encode the gold adsorption sites (*Top*, *H1* and *H2*). Corrections of the weights during the training procedure were performed by the Levenberg–Marquardt algorithm [26,27] and the number of neurons in the hidden layer was optimized. Before the training, the whole training set was randomly partitioned into a learning and validation set, each one with 50% of the objects. Full cross validation of the entire training set was performed by the Leave-One-Out method (LOO). The logistic activation function was used (a sigmoidal) and each input and output variable was linearly normalized between 0.1 and 0.9 on the basis of the training set. The maximum number of iterations in the training was 5000 or 10,000. The training stopped when no further improvement in the root mean squared error (RMSE) for the validation set [28] was observed. After the training,

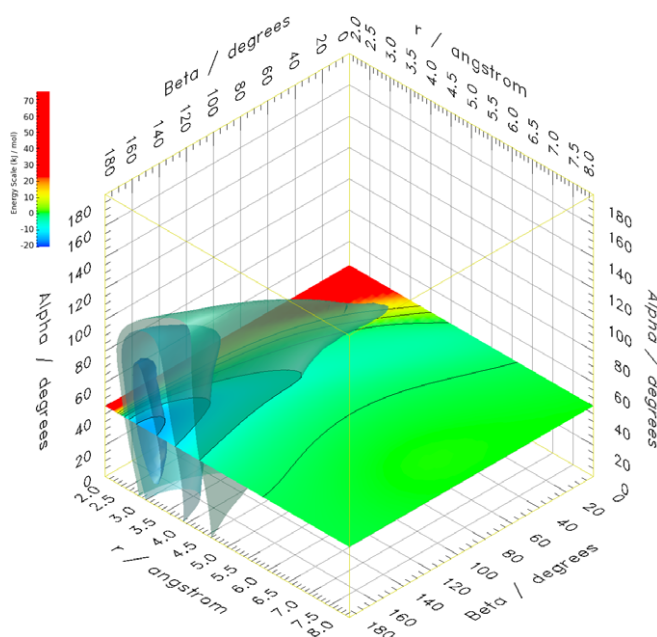


Fig. 6. PES of ethanol over the *H1* site. Three isoennergetic surfaces are displayed at -14 , -10 and -5 kJ mol^{-1} . The plane cuts the surfaces at $\alpha = 45^\circ$. The isoennergetic lines correspond to the energies -14 , -10 , -5 , 0 , 5 and 10 kJ mol^{-1} . The colored bar represents the energy scale.

the results were calculated for the learning set, validation set, LOO method and test sets.

2.3. Ensembles of Feed-Forward Neural Networks

An EnsFFNN consists of several independently trained FFNNs, each one contributing with a single prediction [16,17]. The final prediction for an object, in our case the potential energy, is the average of the outputs from all FFNNs of the ensemble. This methodology smoothes out the random fluctuations in the individual FFNNs predictions. The experiments were carried out with the ASNN program [29].

3. Results and discussion

The impact of the number of hidden neurons, and the size of the ensembles, on the accuracy of the models is analysed in terms of the root mean square error:

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^n [(Y_{\text{calc}} - Y_{\text{exp}})^2]}{n}} \quad (7)$$

where Y_{calc} is the predicted value, Y_{exp} is the target value and n is the number of objects.

The RMSE is calculated for the training set and for the different internal and external test sets. After the optimization of these parameters, a discussion of the NN-PES is presented. The minima of the DFT-calculated potential energy for each molecular orientation and adsorption site are compared with the values from the NNs and from the analytical function at the same distances. The three preferential orientation curves for each site are analysed for the test set. The obtained models are used to generate the 3D representation of the PES for each site. The interpolation ability of different EnsFFNNs is also investigated with EnsFFNNs trained with a different partition of the data and with all data available. The comparison between the DFT data, the predictions of the different EnsFFNNs, and the analytical function are analysed in terms of the RMSE, the correlation coefficient of the predictions relatively to the DFT data, and the mean absolute error (MAE):

$$\text{MAE} = \frac{\sum_{i=1}^n |Y_{\text{calc}} - Y_{\text{exp}}|}{n} \quad (8)$$

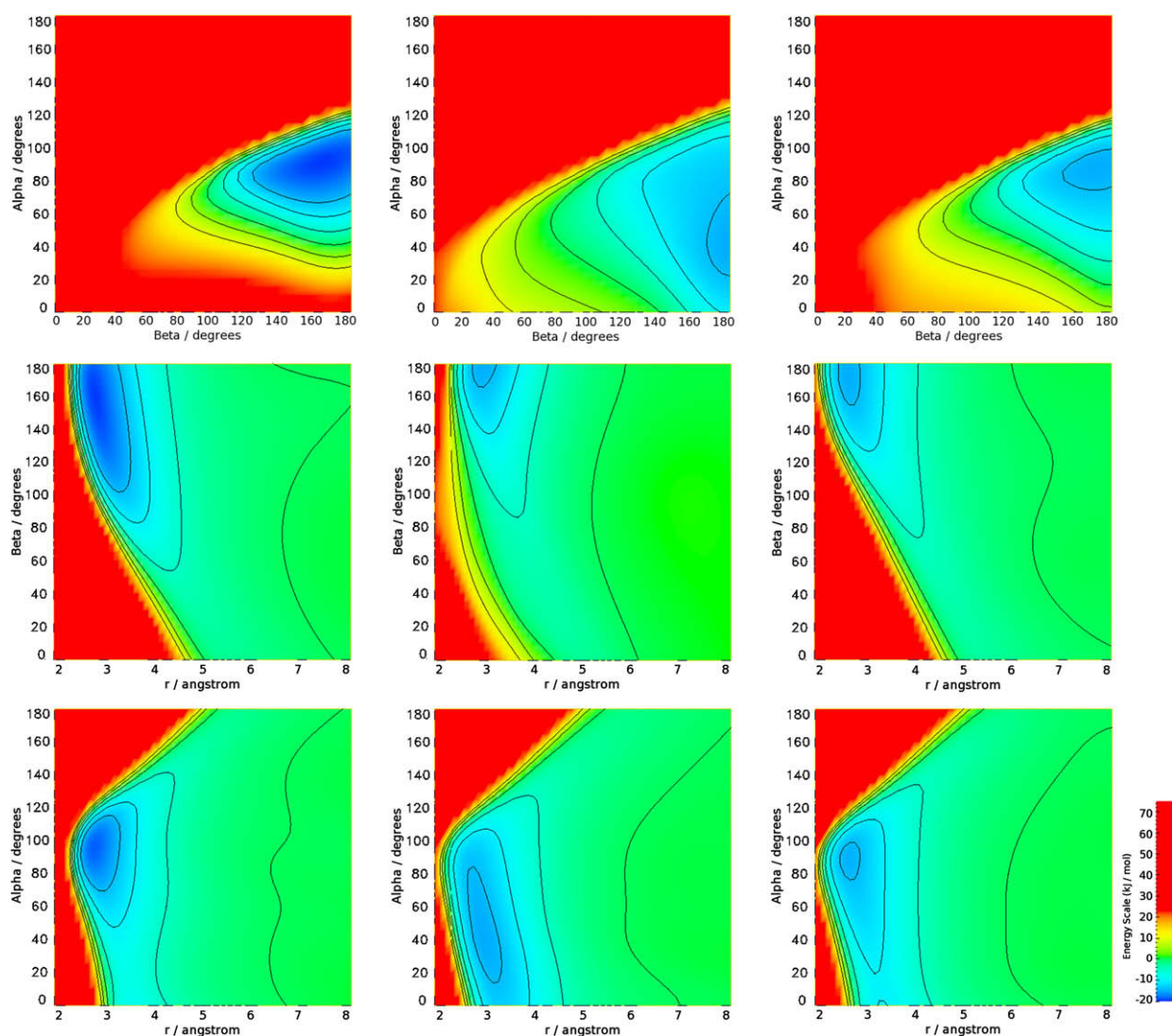


Fig. 7. PES projections for the *Top* (first column), *H1* (second column) and *H2* (third column) sites. First row, from the left to right, projections at $r = 2.8, 3.0, 2.7 \text{ \AA}$; second row, from left to right, projections at $\alpha = 90^\circ, 45^\circ, 90^\circ$; third row, $\beta = 180^\circ$ for all sites. The isoenergetic lines correspond to potential energy of $-14, -10, -5, 0, 5$ and 10 kJ mol^{-1} . The colored bar represents the energy scale.

where Y_{calc} is the predicted value, Y_{exp} is the target value and n is the number of objects.

3.1. Impact of the number of hidden neurons and networks in the ensemble

Ensembles of 15 FFNNs were evaluated for the training, internal validation and external test sets, using different numbers of neurons in the hidden layer: 2–10, 15, 20 and 25. Then EnsFFNNs were trained with different number of networks (1, 5, 10, 15, 20, 25, 50, 75, 100, 150, 200).

Fig. 4 shows a decrease in the RMSE for the test set from $\sim 8 \text{ kJ mol}^{-1}$ for networks with two neurons, to a RMSE of $\sim 5 \text{ kJ mol}^{-1}$ for networks with more than five hidden neurons. The results for the training and other internal validation sets correlate with those for the external test set. Reduction of the RMSE values is more pronounced up to five hidden neurons. Training with more than 10 hidden neurons does not indicate improvements and increases computational requirements. Thus a compromise

of eight hidden neurons has been chosen for the experiments. As for the impact of the ensemble size, the use of ensembles with more than 10 networks has not shown a significant improvement in the predictions.

3.2. Mapping of PES by EnsFFNNs

The minima of the potential energy for each orientation of the ethanol molecule and adsorption sites, from the DFT calculations, and their comparison with the values from the EnsFFNNs and the analytical function at the same distances are presented in Table 1 for the points of the training set.

The EnsFFNNs training results are, in general, in good agreement with those from DFT. Only two cases present absolute errors higher than 1 kJ mol^{-1} : 1.8 kJ mol^{-1} for the 45/180 orientation and 1.1 kJ mol^{-1} for the 90/180 orientation both on the H1 the site.

As for the comparison with the values obtained from the analytical function the EnsFFNNs provide, in general, more accurate predictions for the energy minima. Moreover, it should be emphasized

Table 3
Energy minima from DFT, analytical function and different EnsFFNNs

Ethanol orientation		Distance	Potential energy – U/kJ mol ⁻¹					
α /degrees	β /degrees	O-surface r/Å	DFT	Anal. function	EnsFFNNs ¹	EnsFFNNs ²	EnsFFNNs ³	EnsFFNNs ⁴
<i>Top site</i>								
0	180	3.54	7.76	8.7 (0.94)	7.96 (0.20)	8.68 (0.92)	7.92 (0.16)	8.21 (0.45)
135	180	4.07	5.87	6.7 (0.83)	6.15 (0.28)	5.84 (0.03)	5.78(0.09)	6.02 (0.15)
180	180	6.0	1.37	1.2 (0.17)	1.78 (0.41)	1.47 (0.10)	2.14 (0.77)*	1.57 (0.20)
45	180	3.3	9.04	15.0 (5.96)	9.13 (0.09)	9.68 (0.64)	9.20 (0.16)	9.34 (0.30)
90	0	5.5	2.32	1.5 (0.82)	2.38 (0.06)	1.82 (0.50)	2.31 (0.01)	2.23 (0.09)
90	180	2.67	19.03	19.25 (0.22)	18.17 (0.86)	16.82 (2.21)	17.85(1.18)	17.93 (1.10)
90	90	3.54	11.77	10.4 (1.37)	11.95 (0.18)	11.53 (0.24)	11.81(0.04)	11.25 (0.52)
45	60	3.79	7.0	8.78 (1.78)*	-2.70 (9.70)*	5.15 (1.85)*	5.94(1.06)*	6.83 (0.17)
45	120	3.37	9.81	14.14 (4.33)*	5.65 (4.16)*	12.29 (2.48)*	9.68(0.13)	9.88 (0.07)
75	150	2.81	15.91	18.85 (2.94)*	17.09 (1.18)*	16.89 (0.98)*	15.77(0.14)	15.75 (0.16)
30	180	3.37	7.75	13.36 (5.61)*	8.10 (0.35)*	8.30 (0.55)*	7.82(0.07)*	7.97 (0.22)
60	180	3.17	11.52	16.63 (5.11)*	11.93 (0.41)*	12.41 (0.89)*	11.44(0.08)	11.49 (0.03)
120	180	3.17	11.84	12.13 (0.29)*	11.49 (0.35)*	9.72 (2.12)*	12.28 (0.44)*	11.61 (0.23)
<i>H1 site</i>								
0	180	3.3	13.14	10.7 (2.44)	12.93 (0.21)	13.36 (0.22)	12.92(0.22)	13.15 (0.01)
135	180	4.29	4.34	5.5 (1.16)	3.96 (0.34)	4.11 (0.23)	4.23(0.11)	4.23 (0.11)
180	180	6.04	1.35	0.9 (0.45)	1.51 (0.16)	1.36 (0.01)	3.44(2.09)*	1.38 (0.03)
45	180	2.87	16.42	13.3 (3.12)	14.61 (1.81)	14.81 (1.61)	15.74(0.68)	15.23 (1.19)
90	0	5.57	2.35	1.4 (0.95)	2.46 (0.11)	2.69 (0.34)	2.42 (0.07)	2.37 (0.02)
90	180	2.87	12.80	15.5 (2.7)	13.91 (1.11)	12.93 (0.13)	12.94(0.14)	13.49 (0.69)
90	90	3.87	5.99	7.1 (1.11)	6.03 (0.04)	6.00 (0.01)	5.71 (0.28)	5.68 (0.31)
45	60	3.79	6.69	7.20 (0.51)*	3.07 (3.62)*	6.35 (0.34)*	1.48 (5.21)*	6.45 (0.24)
45	120	3.17	12.50	13.03 (0.53)*	7.48 (5.02)*	11.74 (0.76)*	12.26(0.24)	12.2 (0.30)
75	150	2.81	12.58	15.46 (2.88)*	12.82 (0.32)*	13.99 (1.41)*	12.72(0.14)	12.88 (0.30)
30	180	3.17	15.59	12.76 (2.83)*	14.86 (0.73)*	14.41 (1.18)*	15.10(0.49)*	14.9 (0.69)
60	180	2.81	16.71	14.34 (2.37)*	14.68 (2.03)*	15.22 (1.49)*	15.84(0.87)	15.6 (1.11)
120	180	3.58	6.64	10.43 (3.79)*	6.43 (0.21)*	5.87 (0.77)*	5.72(0.92)*	6.37 (0.27)
<i>H2 site</i>								
0	180	3.3	10.46	9.8 (0.66)	10.32 (0.14)	10.20 (0.26)	10.37(0.09)	10.30 (0.16)
135	180	4.2	5.08	5.3 (0.22)	4.96 (0.12)	5.06 (0.02)	4.94 (0.14)	4.78 (0.30)
180	180	6.04	1.48	1.0 (0.48)	1.42 (0.06)	1.44 (0.04)	3.4(1.92)*	1.45 (0.03)
45	180	3.08	11.82	12.7 (0.88)	11.39 (0.43)	11.72 (0.10)	11.77(0.05)	11.58 (0.24)
90	0	5.37	2.33	1.4 (0.93)	2.37 (0.04)	2.47 (0.14)	2.38 (0.05)	2.43 (0.10)
90	180	2.67	15.07	14.7 (0.37)	14.86 (0.21)	14.67 (0.40)	14.68(0.39)	15.20 (0.13)
90	90	3.87	5.75	7.1 (1.35)	6.03 (0.28)	5.84 (0.09)	5.58 (0.17)	5.74 (0.01)
45	60	4.02	4.44	6.55 (2.11)*	2.11 (2.33)*	6.32 (1.88)*	4.10 (0.34)*	4.50 (0.06)
45	120	2.99	10.61	12.09 (1.48)*	3.96 (6.65)*	7.08 (3.53)*	10.36(0.25)	10.26 (0.35)
75	150	2.81	13.78	14.50 (0.72)*	13.24 (0.54)*	14.00 (0.22)*	13.58(0.20)	13.65 (0.13)
30	180	3.17	10.73	12.19 (1.46)*	10.54 (0.19)*	10.52 (0.21)*	10.75(0.02)*	10.64 (0.09)
60	180	2.81	13.09	13.37 (0.28)*	12.59 (0.50)*	12.79 (0.30)*	12.80(0.29)	12.73 (0.36)
120	180	3.37	8.55	10.61 (2.06)*	7.88 (0.67)*	7.35 (1.20)*	7.15(1.40)*	7.97 (0.58)

EnsFFNNs¹ – training with 366 energy points, tested with 414 energy points (results from Tables 1 and 2); EnsFFNNs² – same training and test set as EnsFFNNs¹, different initial weights; EnsFFNNs³ – training with a new partition in training and test set (training with 536 energy points, test with 244 energy points); EnsFFNNs⁴ – training with all data available.

For all orientations the results from the test set are marked with *. In parenthesis is the absolute error.

that while the analytical function shows an evident discrepancy, relatively to the order of the DFT binding energies, for the referred to 45/180 and 90/180 orientations on the *H1* site, the neural networks predict the right DFT energy order.

Table 2 presents the results for the energy minima from the test set, whose elements, as already said, have not participated in the training. The results show, in general, the same level of accuracy as those from the training set, except for the 45/60 and 45/120 orientations that exhibit considerable errors at all the chosen sites. The value for the 45/60 orientation at the *Top* site is even a non-physical one, since it predicts a positive binding energy.

The bad results for those orientations are presumably due to the fact that the β angle range is not well covered in the training set. From the seven orientations used in it only two have a β angle different of 180: the 90/0 and the 90/90 orientations. To set up models allowing more accurate predictions for all the PES regions it is essential that the training set covers a wider range of possible molecular orientations. This issue will be reanalysed in the following Section 3.3 through the generation of a PES by training with the same set as before but with different initial weights, using a different training and test set and training with all available data without an external test set.

The comparison with the results from the analytical function shows that, except for the 45/60 and 45/120 orientations, the EnsFFNNs provide, in general, more accurate predictions for the energy minima: a maximum absolute error of $\sim 2 \text{ kJ mol}^{-1}$ from the NNs against $\sim 6 \text{ kJ mol}^{-1}$ from the function.

Fig. 5 displays the whole NNs PES, from the test set, for the three most attractive orientations on each site and their comparison with the DFT data, as a complement of Table 2 (that only presents the values of the potential energy minima).

In all presented orientations the most considerable deviations from DFT data are always in the repulsive part of the curves. Despite this deviations this part of the curves is less important if we take into consideration that, at 298 K, kT is only 2.4 kJ mol^{-1} , suggesting that the probability of the high repulsive parts of the PES becoming sampled, during a simulation, is very low.

The 120/180 orientation on the *Top* site presents an average deviation from the DFT data, of $\sim 6 \text{ kJ mol}^{-1}$, at the repulsive part of the curve. For the other two orientations the predictions are in good accordance with DFT data over the entire curves. On the *H1* site, very accurate predictions are obtained for the 30/180 and 75/150 orientations, and for the 60/180 orientation a deviation, of $\sim 2 \text{ kJ mol}^{-1}$, is observed in the attractive part of the curve. For *H2* site the NNs predicted curves are in good agreement with the DFT data.

After the training and testing, 83,509 potential energy points have been predicted by the EnsFFNNs for distances of the ethanol to the surface between 2 and 8 Å, with an interval of 0.1 Å, and for α and β angles between 0 and 180°, with an interval of 5°. Each site is treated separately. Fig. 6 displays an example of PES for the *H1* site in a 3D representation.

The cutting plane at $\alpha = 45^\circ$ includes energy minima for the *H1* site, for example at $\beta = 180^\circ$ and $r = 2.9 \text{ Å}$ (see Table 1). The isoenergetic lines show the energy dependence on the distance and the β angle. The surfaces corresponding to the lines 0, 5 and 10 kJ mol^{-1} are not represented just to avoid a heavy picture.

Fig. 7 illustrates PES projections for minima energy regions. The first, second and third columns correspond to the *Top*, *H1* and *H2* sites, respectively. The first, second and third row correspond to fixing the distances, the α and the β angles, respectively.

The examples of the 3D plots and their projections show the smooth and well-behaved PES predicted by the EnsFFNNs, allowing straightforward energy interpolations. This is a relevant aspect regarding their possible use in Monte Carlo or molecular dynamics simulations. As we shall see in the next subsection, such representations also give a good visualization of the improvements introduced in the PES by changing the training and test sets.

3.3. Different training and test sets

The failure or less accurate predictions of the NNs–PES in a few surface regions, mentioned above, are presumably not due to a limitation of the networks learning but to the available information: a limited data set of potential energy points. In fact, the NNs ability of mapping multidimensional data has already been shown, for the most part of the orientations, by the accurate predictions obtained from an external data set using a limited number of orientations in the training set.

In order to evaluate the impact of the available data in the NNs learning, and consequently in the accuracy of the predicted PES, further experiments have been carried out: (i) using the same training and test sets as before but with different (random) training parameters; (ii) training with a new partition of training and test sets; and (iii) training with all available data for (13 orientations for each site corresponding to 780 energy points) without using an external test set.

The energy minima at each orientation are presented in Table 3 together with the results of the last section.

From Table 3 the following conclusions can be drawn. A different ensemble of FFNNs (EnsFFNNs² trained with the same data set but different random training parameters) may yield better

Table 4

MAE, RMSE and correlation coefficient for learning, validation and test sets for different ensembles and analytical function

	MAE/kJ mol ⁻¹			RMSE/kJ mol ⁻¹			Correlation coefficient		
	Learn	Valid	Test	Learn	Valid	Test	Learn	Valid	Test
<i>For all energy points</i>									
EnsFFNNs ¹	0.21	0.70	2.15	0.38	1.67	4.84	0.9996	0.9915	0.8939
EnsFFNNs ²	0.24	0.67	1.83	0.44	1.78	6.37	0.9994	0.9903	0.8131
EnsFFNNs ³	0.14	0.36	1.83	0.22	0.94	4.94	0.9998	0.9969	0.9047
EnsFFNNs ⁴	0.21	0.45	–	0.52	1.30	–	0.9990	0.9721	–
<i>Only for energy minima</i>									
Anal. function	1.29	–	2.28	1.83	–	2.78	0.9789	–	0.7749
EnsFFNNs ¹	0.35	0.58	2.13	0.54	0.77	3.36	0.9955	0.9917	0.8506
EnsFFNNs ²	0.39	0.44	1.23	0.67	0.76	1.5	0.9939	0.9925	0.9117
EnsFFNNs ³	0.24	0.37	1.23	0.35	0.50	1.83	0.9985	0.9970	0.9035
EnsFFNNs ⁴	0.29	0.43	–	0.42	0.60	–	0.9972	0.9390	–

EnsFFNNs¹ – training with 366 energy points, tested with 414 energy points (results from Tables 1 and 2); EnsFFNNs² – same training and test set as EnsFFNNs¹, different initial weights; EnsFFNNs³ – training with a new partition in training and test set (training with 536 energy points, test with 244 energy points); EnsFFNNs⁴ – training with all data available.

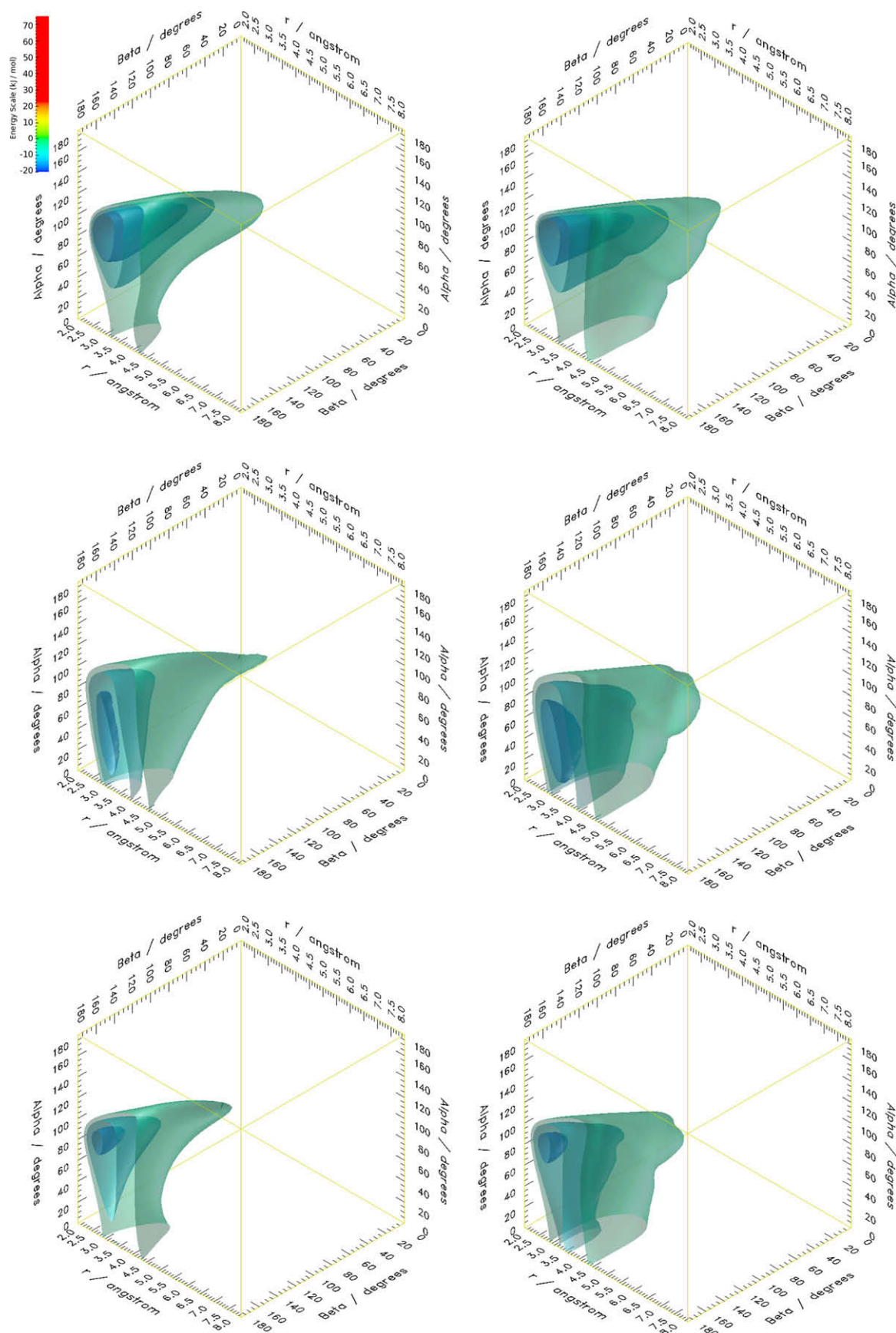


Fig. 8. PES from training with the first set (first column) and with all available data (second column) for the *Top* (first row), *H1* (second row) and *H2* (third row) sites. Isoenergetic surfaces at: -14 , -10 and -5 kJ mol $^{-1}$. The colored bar represents the energy scale. (For interpretation of the references in color in this figure legend, the reader is referred to the web version of this article.)

predictions for some orientations, namely 45/60 and 45/120, but the performance degrades for other tests (the overall RMSE for the validation and test sets) (see Table 4). However, if it is only considered the predictions for the energy minima of each orientation the EnsFFNNs² results are more accurate: a RMSE of 1.5 kJ mol⁻¹ against 3.36 kJ mol⁻¹ for the test set. Also, the correlation coefficients between the DFT results for the energy minima and the EnsFFNNs² are improved from 0.85 to 0.91 (Table 4).

A similar situation was observed by using a different partition of the data set into training and test sets. The RMSE are improved from 3.36 kJ mol⁻¹ to 1.83 kJ mol⁻¹ for the test set (Table 4). The new orientations used in the training help the NNs to learn much better the interactions at 45/60 and 45/120 orientations. Yet, the binding energies at orientations 180/180 in the sites *H1* and *H2* and 45/60 in *H1* site are less accurately predicted. These results indicate that the training sets did not cover well enough the configuration space, leading to some instability in the results, as well as large errors for some orientations. Even though, excellent predictions were achieved for most of the orientations in the independent test set.

Fig. 8 shows isoenergetic surfaces for the three sites trained with the initial set and with all available data.

The isoenergetic surfaces allow the visualization of the improvement for the regions not so well represented with the first training set. When the training uses all data available, the surfaces cover, in general, wider regions in the configurational space. As already referred to before, it is very important to cover as much as possible the configurational space in the training in order to obtain accurate predictions.

In the course of the present experiments, Associative Neural Networks have also been probed. An ASNN is a combination of a memory-less (after the training all information about the input patterns are stored in the NN weights of the networks) and a memory-based method (the data used to build the models are also stored in a “memory” and the predictions are corrected based on some local approximations of the stored examples). The results obtained by ASNNs are not presented and discussed here because they are similar to the ones from EnsFFNNs. Nonetheless, this is an important methodology to take into account – it allows to incorporate new data in the memory after the training is finished, making it possible to improve predictions with new data without the need to retrain the NNs. It has a potential application in the course of our future experiments, when DFT data for new orientations and sites will become gradually available and kept in the “memory” of the system.

4. Conclusions

The present results indicate that NNs can be trained to map PES with suitable accuracy to be used in molecular simulations, particularly for the ethanol–Au(111) interactions. Once the networks are well trained they are able to produce, as output, any required number of energy points for numerical interpolations, with similar or better accuracy than other mapping methods.

EnsFFNNs give better results than single FFNNs, showing their capability of taking into account the most subtle features of this type of interactions when the model was tested for orientations that did not participate in the learning procedure.

The NNs–PES have to be tested in molecular simulations. The test will be carried out using the tabular potential energies, predicted by the NNs, for working out thermal, structural and dynamical properties to be compared with the preliminary Monte Carlo simulation values already obtained from the analytical function [1]. One point has already turned out, however, in the present work: in general, the NNs can reproduce the DFT results with a bet-

ter accuracy than the analytical function as far as the probed interaction sites are concerned.

Work is in progress regarding a much finer screening, by DFT, of the different gold interaction sites and ethanol orientations. Indeed, before a full simulation test can be performed, based on the NNs data, it is necessary to obtain DFT results for sites other than the *Top*, *H1* and *H2*. Such results will certainly increase the accuracy of the NNs mappings, using different memories as the new data is becoming available.

Finally, it should be mentioned that the representation of metallic surfaces by cluster models has been a common approximation in order to minimise the heavy computational requirements. We have commented on that elsewhere [1,30]. Yet, nowadays, accurate periodic DFT methods can be implemented in relatively low cost computer networks. We will consider them in future applications to surface models. Nevertheless, the conclusions of the present paper do not depend on the choice of the system and model but rather indicate that NNs offer an alternative to be used in Monte Carlo and molecular dynamics simulations. As for molecular dynamics, which usually require continuously differentiable potentials, the interpolation routines may, however, slow down the simulations.

Acknowledgements

The authors acknowledge Fundação para a Ciência e Tecnologia (FCT, Lisbon, Portugal) for financial support, and for a Ph.D. grant (SFRH/BD/18347) awarded to D.A.R.S. Latino. The authors also thank Dr. Igor Tetko who made available the software to implement EnsFFNNs and ASNNs.

References

- [1] R.P.S. Fartaria, F.F.M. Freitas, F.M.S.S. Fernandes, *Int. J. Quantum Chem.* 107 (11) (2007) 2169.
- [2] M.Y. Ballester, A.J.C. Varandas, *Phys. Chem. Chem. Phys.* 7 (2005) 2305.
- [3] J.N. Murrell, S. Carter, S.C. Farantos, P. Huxley, A.J.C. Varandas, *Molecular Potential Energy Functions*, John Wiley & Sons, London, 1984.
- [4] K.T. No, B.H. Chang, S.Y. Kim, M.S. Ihon, H.A. Scheraga, *Chem. Phys. Lett.* 271 (1997) 152.
- [5] F.V. Prudente, J.J.S. Neto, *Chem. Phys. Lett.* 287 (1998) 585.
- [6] A.C.P. Bittencourt, F.V. Prudente, J.D.M. Vianna, *Chem. Phys.* 297 (2004) 153.
- [7] T.M.R. Filho, Z.T. Oliveira, L.A.C. Malbouissou, R. Gargano, J.J.S. Neto, *Int. J. Quantum Chem.* 95 (2003) 281.
- [8] H. Gassner, M. Probst, A. Lauenstein, K. Hermansson, *J. Phys. Chem. A* 102 (1998) 4596.
- [9] K.-H. Cho, K.T. No, H.A. Scheraga, *J. Mol. Struct.* 641 (2002) 77.
- [10] J.B. Witkoskie, D.J. Doren, *J. Chem. Theory Comput.* 1 (2005) 14.
- [11] D.A.R.S. Latino, F.F.M. Freitas, J. Aires-de-Sousa, F.M.S.S. Fernandes, *Int. J. Quantum Chem.* 107 (2007) 2120.
- [12] S. Lorenz, A. Gross, M. Sheffler, *Chem. Phys. Lett.* 395 (2004) 210.
- [13] S. Lorenz, M. Sheffler, A. Gross, *Phys. Rev. B* 73 (2006) 115431.
- [14] S. Manzhos, T. Carrington Jr., *J. Chem. Phys.* 125 (2006) 084109.
- [15] L.M. Ralf, M. Malshe, M. Hagan, D.I. Doughan, M.G. Rockley, R. Komanduri, *J. Chem. Phys.* 122 (2005) 084104.
- [16] D.K. Agrafiotis, W. Cedeno, V.S. Lobanov, *J. Chem. Inf. Comput. Sci.* 42 (2002) 903.
- [17] T.G. Dietterich, *Ensemble Learning*, in: *The Handbook of Brain Theory and Neural Networks*, Arbib, MIT Press, Cambridge, MA, 2002, pp. 405–408.
- [18] A.D. Becke, *J. Chem. Phys.* 98 (1993) 5648.
- [19] A.D. Becke, *J. Chem. Phys.* 98 (1993) 1372.
- [20] P.J. Hay, W.R. Wadt, *J. Chem. Phys.* 82 (1985) 270.
- [21] W.J. Hehre, R. Ditchfield, J.A. Pople, *J. Chem. Phys.* 56 (1972) 2257.
- [22] M.J. Frisch, G.W. Trucks, H.B. Schlegel, G.E. Scuseria, M.A. Robb, J.R. Cheeseman, V.G. Zakrzewski, J.A. Montgomery, R.E. Stratmann, J.C. Burant, S. Dapprich, J.M. Millam, A.D. Daniels, K.N. Kudin, M.C. Strain, O. Farkas, J. Tomasi, V. Barone, M. Cossi, R. Cammi, B. Mennucci, C. Pomelli, C. Adamo, S. Clifford, J. Ochterski, G.A. Petersson, P.Y. Ayala, Q. Cui, K. Morokuma, D.K. Malick, A.D. Rabuck, K. Raghavachari, J.B. Foresman, J. Cioslowski, J.V. Ortiz, B.B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R. Gomperts, R.L. Martin, D.J. Fox, T. Keith, M.A. Al-Laham, C.Y. Peng, A. Nanayakkara, C. Gonzalez, M. Challacombe, P.M.W. Gill, B.G. Johnson, W. Chen, M.W. Wong, J.L. Andres, M. Head-Gordon, E.S. Replogle, J.A. Pople, Gaussian 98, Gaussian Inc., Pittsburgh PA, 1998.
- [23] J. Zupan, J. Gasteiger, *Neural Networks in Chemistry and Drug Design*, Wiley-VCH, Weinheim, 1999.
- [24] I.V. Tetko, *J. Chem. Inf. Comput. Sci.* 42 (2002) 717.

- [25] I.V. Tetko, *Neural Process. Lett.* 16 (2002) 187.
- [26] W.H. Press, S.A. Teukolsky, W.T. Vetterlung, B.P. Flannery, *Numerical Recipes in C*, second ed., Cambridge University Press, New York, 1994. p. 998.
- [27] A.J. Shepherd, *Second Order Methods for Neural Networks*, Springer-Verlag, London, 1997. p. 145.
- [28] M. Bishop, *Neural Networks for Pattern Recognition*, Oxford University Press, Oxford, 1995.
- [29] VCCLAB, Virtual Computational Chemistry Laboratory, 2005. <<http://www.vcclab.org>>.
- [30] R.S. Neves, A.J. Motheo, R.P.S. Fartaria, F.M.S.S. Fernandes, *J. Electroanal. Chem.* 609 (2007) 140.