Supporting Information

Chemically Interpretable Machine Learning for Predicting HER Activity in Au-Based Alloys

Maja Kubik, Shiqi Wang*, Pedro H. C. Camargo*

University of Helsinki, Department of Chemistry, A.I. Virtasen aukio 1, Helsinki, Finland

**Corresponding authors. Email: shiqi.z.wang@helsinki.fi and pedro.camargo@helsinki.fi*

Equations (Descriptor Definitions)

Equation S1. Generalized coordination number (GCN), defined as a continuous measure of local coordination relative to the bulk maximum. GCN distinguishes adsorption motifs (ontop, bridge, hollow) and serves as a geometric descriptor for hydrogen binding.

$$GCN = \frac{\sum_{i=1}^{N} CN_i}{CN_{max}}$$

where CN_i is the coordination number of the i[th] surface atom, CN_{max} is the maximum bulk coordination number and N is the number of atoms at the adsorption site.

Equation S2. Unit cell volume per atom (Å^3), calculated from the stoichiometry-weighted volumes of pure elements A and B. This descriptor provides a proxy for bond length trends in alloy systems.

$$V_{alloy/atom} = \frac{nV_A + mV_B}{n + m}$$

where V_A and V_B are the unit-cell volumes of pure elements A and B, respectively, retrieved from MP. For each unique element MP was queried for the lowest-energy cubic polymorph.

Equation S3. Weighted ionization energy (WIE, eV), defined as the stoichiometry-weighted average of the elemental ionization energies of A and B. This descriptor reflects the strength with which valence electrons are bound.

$$WIE = \frac{(n \times IE_A + m \times IE_B)}{(n + m)}$$

where IE_A and IE_B are ionization energies of element A and B respectively, and n and m are their stoichiometric indices.

Equation S4. Weighted electronegativity (WEN, Pauling), calculated as the stoichiometry-weighted average of elemental electronegativities of A and B. This descriptor captures the ability of alloy sites to attract charge.

$$WEN = \frac{(n \times EN_A + m \times EN_B)}{(n + m)}$$

where EN_A and EN_B are Pauling electronegativities of element A and B respectively, and n and m are their stoichiometric indices.

Equation S5. Weighted atomic radius (WAR), the stoichiometry-weighted average of elemental radii of A and B. This descriptor provides a size-based measure that correlates with orbital overlap and alloy geometry.

$$WAR = \frac{(n \times r_A + m \times r_B)}{(n + m)}$$

where r_A and r_B are atomic radii of element A and B respectively, and n and m are their stoichiometric indices.

Equation S6. Site-specific electronic descriptor (Ψ), integrating the outer valence electron count and Pauling electronegativity of atoms in the first coordination shell. Ψ encodes the immediate electronic environment at the adsorption site and strongly influences hydrogen binding.

$$\psi = \frac{(\prod_{i=1}^{N} S_i)^{\frac{2}{N}}}{(\prod_{i=1}^{N} EN_i)^{\frac{1}{N}}}$$

where S_i is the outer valence electron count and EN_i the Pauling electronegativity of the i^{th} atom at the adsorption site, and N is the total count of first-neighbour atoms.

Tables (Dataset and Model Details)

Table S1. Mapping of coefficients used to calculate the site-specific electronic descriptor (Ψ). For each adsorption motif, the normalized site label, stoichiometry, and number of A- and B-type atoms (nA, nB) in the first coordination shell are listed, together with the total number of neighbors (N). These coefficients define the local atomic environments considered in the construction of Ψ.

Normalised site label	Stoichiometry	n of A	n of B	N
bridge-tilt AA, bridge AA	A3B, AB3, AB, A	2	0	2
bridge-tilt AB, bridge AB	A3B, AB3, AB	1	1	2
bridge-tilt BB, bridge BB	A3B, AB3, AB	0	2	2
fcc AAA, hollow-tilt AAA fcc	A3B, AB3, AB, A	6	0	6
fcc AAB, hollow-tilt AAB fcc	A3B, AB3, AB	4	2	6
fcc ABB, hollow-tilt ABB fcc	A3B, AB3, AB	2	4	6
hcp AAA, hollow-tilt AAA hcp	A3B, AB3, AB, A	3	0	3
hcp AAB, hollow-tilt AAB hcp	A3B, AB3, AB	2	1	3
hcp ABB, hollow-tilt ABB hcp	A3B, AB3, AB	1	2	3
top-tilt A, top A	A3B, AB3, AB	4	6	10
top-tilt A, top A	A	10	0	10
top-tilt B, top B	A3B, AB3, AB	6	4	10

Table S2. Values of generalized coordination number (GCN) for canonical adsorption sites (fcc, hcp, bridge, ontop) on fcc(111) and fcc(101) surfaces, adapted from Martínez-Alonso et al. (2024) [43]. These GCN values provide the geometric basis for encoding coordination environments in the feature set.

Surface	Geometry	Site	*GCN*
111	fcc	fcc	5.25
111	fcc	hcp	3.25
111	fcc	ontop	0.75
111	fcc	bridge	0.00
101	fcc	fcc	5.25
101	fcc	hcp	3.25
101	fcc	ontop	0.75
101	fcc	bridge	0.00

Table S3. Per-fold mean absolute error (MAE, eV) for Extra Trees (ET), Random Forest (RF), and XGBoost models under the 5-fold StratifiedGroupKFold split. Results are reported for the training set and for the Au-containing subset, with mean ± standard deviation across folds listed in the bottom row.

Fold	ET		RF		XGBoost	
	MAE	MAE_{Au}	MAE	MAE_{Au}	MAE	MAE_{Au}
K = 1	0.1432	0.1486	0.1604	0.1682	0.1587	0.1595
K = 2	0.1473	0.1744	0.1635	0.1925	0.1619	0.1788
K = 3	0.1478	0.2089	0.1619	0.2218	0.1569	0.2126
K = 4	0.1414	0.1299	0.1574	0.1522	0.1625	0.1625
K = 5	0.1244	0.1006	0.1444	0.1321	0.1497	0.1497
Mean ± SD	0.1408 ± 0.0095	0.1525 ± 0.0415	0.1575 ± 0.0077	0.1733 ± 0.0350	0.1579 ± 0.0051	0.1724 ± 0.0247

Table S4. Optimized hyperparameters for Extra Trees, Random Forest, and XGBoost models, together with the ranges explored during grid search. Best values were selected using the group-aware validation protocol.

Model	Parameter	Best value	Search space
Extra Trees	n_estimators	1000	[1000]
	max_features	9	[1, 2, 3, 4, 5, 6, 7, 8, 9]
	max_depth	100	[100, 300, 600, 900]
	min_sample_split	2	[2, 3, 4]
Random Forest	n_estimators	1000	[1000]
	max_features	6	[1, 2, 3, 4, 5, 6, 7, 8, 9]
	max_depth	100	[100, 300, 600, 900]
	min_sample_split	2	[2, 3, 4]
XGBoost	n_estimators	800	[200, 400, 600, 800]
	learning_rate	0.075	[0.05, 0.075, 0.10, 0.15]
	max_depth	6	[3, 4, 5, 6]
	subsample	0.8	[0.6, 0.8, 1.0]
	colsample_bytree	1	[0.6, 0.8, 1.0]
	min_child_weight	1	[1, 3, 5]
	gamma	0	[0, 0.1, 0.2]
	reg_alpha	0.0001	[0, 0.0001, 0.001]
	reg_lambda	1	[0.5, 1, 2]

Table S5. Site-resolved test metrics for the unified Extra Trees model. For each adsorption site type, the table reports sample sizes (n, nAu), mean absolute error (MAE, eV), root-mean-square error (RMSE, eV), and coefficient of determination (R^2). Errors are lowest for close-packed sites (fcc, hcp) and highest for ontop motifs, consistent with site stability trends.

Site type	n	n_{Au}	MAE	MAE_{Au}	R^2	R^2_{Au}
ontop	321	26	0.2493	0.2216	0.4548	0.3148
bridge	312	20	0.1560	0.1661	0.8104	0.5841
fcc	393	20	0.0658	0.0538	0.9612	0.9764
hcp	338	24	0.0719	0.0863	0.9626	0.9405

Table S6. Performance of site-specific models trained separately on ontop, bridge, fcc, and hcp subsets. For each site type, sample sizes (n, nAu) and test metrics (MAE, RMSE, R^2) are reported. Despite specialization, these models did not outperform the unified, site-aware Extra Trees model; results are provided for completeness.

Site type	**n**	**n_{Au}**	MAE	MAE_{Au}	R^2	R^2_{Au}
ontop	321	26	0.2530	0.2304	0.4433	0.2235
bridge	312	20	0.1607	0.1775	0.7988	0.5381
fcc	393	20	0.0848	0.0819	0.9374	0.9541
hcp	338	24	0.1090	0.1100	0.8939	0.9128
combined	1394	90	0.1478	0.1535	0.8362	0.7927

Table S7. Canonical adsorption site types used in this study (ontop, bridge, fcc, hcp) and the corresponding normalized site-label variants grouped under each type. This mapping defines the consistent site categories employed in model training and analysis.

Site type	Site variants
ontop	ontop A, ontopB
bridge	bridgeAA, bridgeBB, bridgeAB
fcc	fccAAA, fccBBB, fccABB, fccAAB
hcp	hcpAAA, hcpBBB, hcpABB, hcpAAB

Figures

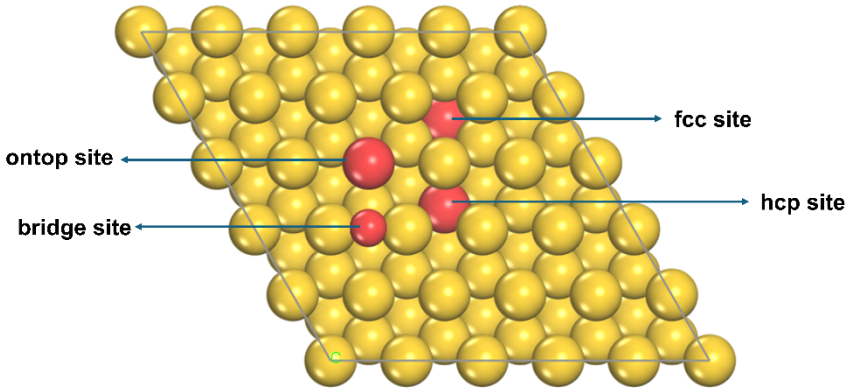

Figure S1. Schematic illustration of the four characteristic adsorption site types, ontop, bridge, fcc hollow, and hcp hollow, on an fcc (111) surface. Red spheres indicate hydrogen adsorption positions, while gold spheres represent surface metal atoms.

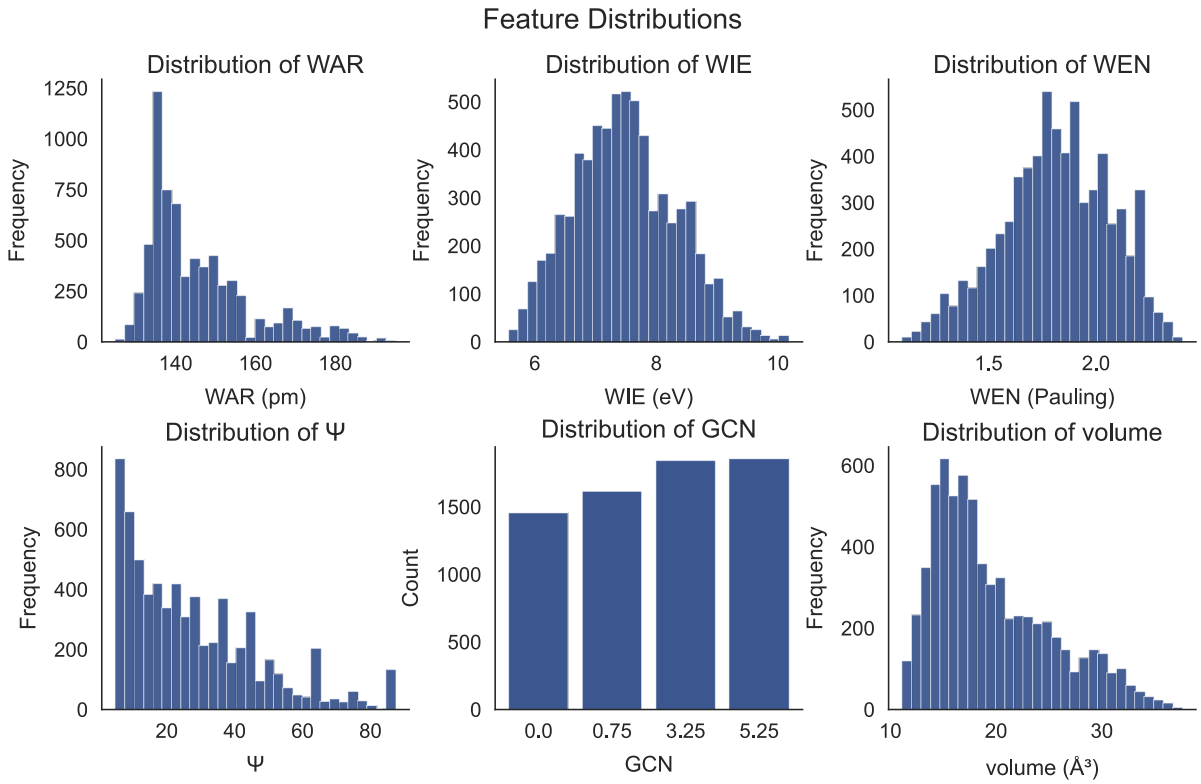

Figure S2. Histograms of engineered descriptors in the HER-only training set (N = 1,394). Distributions are shown for weighted atomic radius (WAR, pm), weighted ionization energy (WIE, eV), weighted electronegativity (WEN, Pauling), site-specific electronic descriptor (Ψ, dimensionless), generalized coordination number (GCN, discrete), and unit-cell volume per atom (Å³). These plots confirm the chemical diversity of the curated dataset and the broad coverage of geometric and electronic environments.

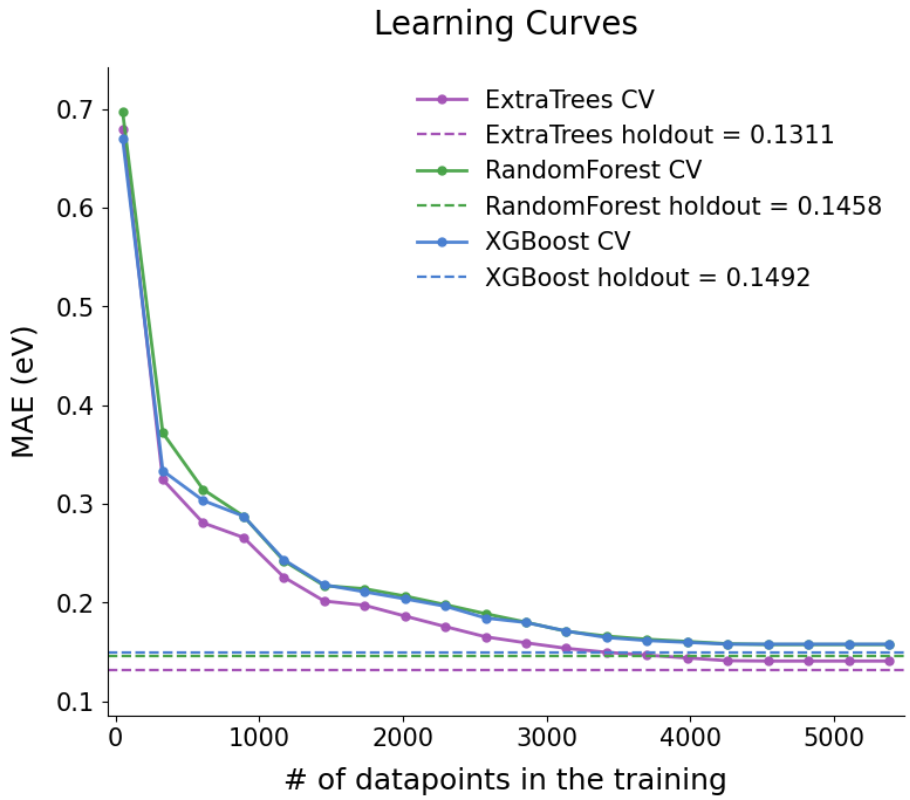

Figure S3. Learning curve analysis for Extra Trees, Random Forest, and XGBoost models. MAE (eV) is plotted against training set size using final hyperparameters. Dashed horizontal lines mark the holdout (test set) MAE. These diagnostics confirm the stability of training and the convergence of model accuracy with increasing data. The plateau beyond ~3,500 training datapoints suggests that the feature potential has been largely exhausted, and further improvements are unlikely without more informative descriptors.

DFT Methodology

DFT calculations were performed with the first-principles simulation Cambridge Sequential Total Energy Package (CASTEP) module in Materials Studio software. The exchange-correlation potential was described by the generalized gradient approximation (GGA) with the Perdewe-Burkee-Ernzerhof (PBE) functional. The interactions between valence electrons and ionic cores were described by the OTFG ultrasoft pseudo-potential method. A plane-wave basis set with a cutoff energy of 400 eV was assigned to the potential method. The empirical dispersion correction in Grimme's scheme was employed to consider the van der Waals (vdW) interaction. The Broyden-Fletcher-Goldfarb-Shannon (BFGS) algorithm with a medium quality setting of k-points was used for all the energy minimizations in this work. The Au-M model is constructed by a 6-layerd 3×3×3 supercell with 108 atoms. The geometry optimization convergence tolerances for the energy change, maximum force and maximum displacement were 5×10^{-5} eV/atom, 0.001 eV/Å, and 0.005 Å, respectively.