ESI for Machine-learning based prediction of small molecule – surface interaction potentials

Ian Rouse[*a]   Vladimir Lobaskin [a]

[a] School of Physics, University College Dublin, Belfield, Dublin 4, Ireland
[*] ian.rouse@ucd.ie

December 20, 2022

1 Additional information

In Table S1 we provide SMILES codes and ID tags for all the molecules used in the development of the PMF prediction model, with the additional chemicals generated by selection of highly-cited SMILES codes from ChemSpider and miscellanous small fragments in S2 and S3. Structures for all chemicals and their ID tags are shown in Figure S1, with chemicals sharing a SMILES code grouped together into a single figure. These SMILES codes may be mapped to multiple 3D structures or parameterisations if the same SMILES code is used for multiple epimers, enantiomers, etc or if the same molecule is parameterised using both GAFF and CHARMM. We assign a unique ID to each structure based on their equivalent amino acid for side chain analogues, common name or ChemSpider ID and using the suffix "-AC" to denote GAFF parameterisations or "-JS" for CHARMM. In Table S4 we list details of the surfaces considered in this work, including additional ones parameterised for the generation of PMFs for structures not yet covered by metadynamics. Table S5 presents the R^2, Pearson correlation, and mean KL divergence for each of the fifteen trained models using the final version of the methodology to identify which models successfully predict PMFs outside of their training set. The computed binding energies for the class of novel chemicals on FCC metal surfaces are provided in Table S6. To illustrate the predictive power of the model for the biochemical set of interest for UnitedAtom, we provide example tables of the predicted adsorption energies of the training molecules to amorphous carbon in Table S7 and the Cu (111) surface in Table S8. Figure S2 presents a comparison between the predictions for the testing chemical BGALNA on nine metal surfaces, including three copper surfaces not included in the training set and compares these to the closely related molecule BGLCNA and gives examples of the input potentials for these surfaces. In Figures S3 - S6 we show predicted potentials of mean force for a selection of four surfaces, comparing the input PMFs to the predictions generated using parameters matching the particular set of metadynamics used for their input and the canonical form using a set of consistent parameters and surface location determined from the potential probes for all predictions. We show these predictions for the gold (100) and CdSe surfaces to demonstrate strongly and weakly binding planar surfaces and to a hydrated TiO_2 surface and surface-modified CNT to demonstrate more complex planar and cylindrical surfaces respectively. In these figures, the adsorbates are labelled omitting the SCA suffix for consistency with the UnitedAtom[11] naming convention.

ID	Source(s)	SMILES	Description
ALASCA	All	C	Alanine side chain (methane)
ARGSCA	All	CCCNC(N)=[NH2+]	Arginine side chain (1-Propylguanidinium)
ASNSCA	All	CC(N)=O	ASN side chain (acetamide)
ASPSCA	All	CC(=O)[O-]	ASP side chain (acetate)
CYSSCA	All	CS	CYS side chain (carbon sulfide)
GLNSCA	All	CCC(N)=O	GLN side chain (propionamide)
GLUSCA	All	CCC(=O)[O-]	GLU side chain (propionate)
HIDSCA	All	Cc1cnc[nH]1	HID side chain (δ-4-Methylimidazole)
HIESCA	All	Cc1c[nH]cn1	HIE side chain (ε-4-Methylimidazole)
ILESCA	All	CCCC	ILE side chain (butane)
LEUSCA	All	CC(C)C	LEU side chain (isobutane)
LYSSCA	All	CCCC[NH3+]	LYS side chain (butylammonium)
METSCA	All	CCSC	MET side chain (ethyl methyl sulfide)
PHESCA	All	Cc1ccccc1	PHE side chain (toluene)
SERSCA	All	CO	SER side chain (methanol)
THRSCA	All	CCO	THR side chain (ethanol)
TRPSCA	All	Cc1c[nH]c2ccccc12	TRP side chain (3-methylindole)
TYRSCA	All	Cc1ccc(O)cc1	TYR side chain (p-cresol)
VALSCA	All	CCC	VAL side chain (propane)
HIPSCA,HSPSCA	All but SU-1	Cc1c[nH]c[nH+]1	Protonated HIS side chain
PROSCA	UCD	C1CC1	Cyclopropane
GANSCA, GLUPSCA	All but SU-1	CCC(=O)O	Protonated GLU side chain
ASPPSCA	UCD	CC(=O)O	Protonated ASP side chain
CYM	Both	C[S-]	Charged cysteine side chain
GLY	SU-2	NCC(=O)O	Glycine
PRO	SU-2	O=C(O)C1CCCN1	Proline
ETA, MAMM	All but SU-1	C[NH3+]	Ethanolamine
DMEP,PHO	All but SU-1	COP(=O)([O-])OC	Phosphate
CHL, NC4	All but SU-1	C[N+](C)(C)C	Tetramethylammonium
EST,MAS	All but SU-1	COC(C)=O	Ester linkage
AFUC	UCD	CC1OC(O)C(O)C(O)C1O	Alpha fucose
BGLCNA	UCD	CC(=O)NC1C(O)OC(CO)C(O)C1O	2-acetyl-2-deoxy-beta-d-glucosamine
AMAN	UCD	OCC1OC(O)C(O)C(O)C1O	Alpha-d-mannose
DGL, BGLC	All but SU-1	OCC1OC(O)C(O)C(O)C1O	Beta-d-glucose
BGALNA	UCD	CC(=O)NC1C(O)OC(CO)C(O)C1O	2-acetyl-2-deoxy-beta-d-galactoseamine
CHOL	UCD	C[N+](C)(C)CCO	Choline

Table S1 A summary of the chemicals present in the dataset used to build the model for the prediction of PMFs, containing the ID(s) for that molecule used, which PMF computational methods have been used for molecule, the SMILES code, formal charge, and a comment describing the chemical in question and providing the equivalent chemical for the side-chain analogues. In the toolkit, the ID is modified by "-JS" or "-AC" to distinguish between structures provided by UCD or generated using Charmm GUI and generated by ACPype respectively.

ID	Smiles	Description
ETHANE	CC	Ethane
PROPANE	CCC	Propane
ETHENE	C=C	Ethene
PROPENE	C=CC	Propene
BUTENE1	C=CCC	But-1-ene
BUTENE2	CC=CC	But-2-ene
BUTENE13	C=CC=C	But-1,3-ene
CS-3	CC(CN)O	1-amino-2-propanol
CS-173	CC(=N)O	Acetamide
CS-210	CC(=O)CN	Aminoacetone
CS-218	[NH4+]	Ammonium ion
CS-234	C(CN)C(=O)O	Beta-alanine
CS-270	C(=O)(N)[O-]	Carbamate
CS-271	C(=O)(N)O	Carbamic acid
CS-280	C(C(=O)O)ON	(Aminooxy)acetic acid
CS-387	CC(C)[N+](=O)[O-]	2-nitropropane
CS-582	CC(C(=O)O)N	DL-alanine
CS-693	C(=O)N	Formamide
CS-730	C(C(=O)O)N	Glycine
CS-949	C(=O)(C(=O)O)N	Oxamate
CS-1057	CNCC(=O)O	Sarcosine
CS-1113	C[N+](C)(C)[O-]	Trimethylamine oxide
CS-1143	C(=N)(N)O	Urea
CS-1913	C/C(=N/O)/O	Acetohydroxamic acid
CS-3320	c1cc(oc1)CN	Furfurylamine
CS-3530	C(=N)(NO)O	Hydroxyurea
CS-3970	CON	Methoxyamine
CS-5008	C(=N)(NN)O	Semicarbazide
CS-5439	CCOC(=O)N	Urethane
CS-5735	C[C@@H](C(=O)O)N	L-alanine
CS-5894	CN(C)N=O	Dimethylnitrosamine
CS-5993	CN(C)C=O	N,N-dimethylformamide
CS-6110	C=NO	Formoxime
CS-6135	C[N+](=O)[O-]	Nitromethane
CS-6330	CCC(=O)N	Propionamide
CS-6331	C=CC(=O)N	Acrylamide
CS-6332	C(C(=O)N)Cl	2-chloroacetamide
CS-6334	CC(=O)NC	N-methylacetamide
CS-6338	CC[N+](=O)[O-]	Nitroethane
CS-7021	CNC(=O)NC	N,N'-dimethylurea
CS-7610	C(C#N)C(=O)N	2-cyanoacetamide
CS-7683	c1cc(cnc1)O	3-Pyridinol
CS-7720	C(CO)C#N	Hydracrylonitrile
CS-8142	C1CNC(=O)N1	2-imidazolidinone
CS-8537	c1ccnc(c1)O	2-pyridone
CS-8733	C(CN)CO	3-amino-1-propanol
CS-8897	c1cnoc1	Isoxazole
CS-8898	c1cocn1	Oxazole
CS-8953	Cc1cc(on1)C	3,5-Dimethylisoxazole
CS-9357	C(C#N)C(=O)O	Cyanoacetic acid
CS-10471	C1=CC(=O)N=C1O	Maleimide
CS-10955	O=C1CCC(=O)N1	Succinimide
CS-11227	CNC(=N)O	1-Methylurea
CS-11244	CN(C)C(=O)N	1,1-dimethylurea
CS-11530	C1CC(=NC1)O	2-Pyrrolidone

Table S2 Additional chemicals generated via their SMILES code using ACPYPE and parameterised for use in the PMF Prediction toolkit, part 1

ID	Smiles	Description
CS-11727	CN=C=O	Methyl isocyanate
CS-11787	c1cnccc1O	4-pyridone
CS-4945	CC(CO)N	DL-Alaninol
CS-12139	C(=N)(N)NN=O	1-Nitrosoguanidine
CS-12144	C1CCN=C(C1)O	2-Piperidone
CS-12229	c1cc[n+](cc1)[O-]	pyridine oxide
CS-12814	CN1CCCC1=O	1-methyl-2-pyrrolidinone
CS-13254	c1cc([nH]c1)C=O	2-Formyl-1H-pyrrole
CS-13420	C/C(=N/N)/O	Acethydrazide
CS-16629	CN(C)C(=O)C=C	N,N-dimethylacrylamide
CS-19489	c1cncnc1O	1H-Pyrimidin-4-one
CS-20788	CCC(CO)N	(+/-)-2-amino-1-butanol
CS-23329	c1cnc([nH]1)C=O	Imidazole-2-carbaldehyde
CS-28994	CNC=O	N-methylformamide
CS-29107	CC(=O)N(C)C	N,N-dimethylacetamide
CS-61461	c1cc(=O)[nH]nc1	Pyridazinone
CS-61465	C1COC=N1	Oxazoline
CS-61686	c1cnc(nc1)O	Pyrimidone
CS-61707	CC(C)C(=O)N	Isobutyramide
CS-62242	C(C(=N)O)N	Glycinamide
CS-64234	C[C@H](C(=O)O)N	D-(-)-Alanine
CS-65596	c1cnc(cn1)O	Pyrazinol
CS-66578	C(=O)(NN)NN	Carbohydrazide
CS-66579	C1COC(=N1)O	Oxazolidinone
CS-68900	c1c(nc[nH]1)C=O	4-Imidazolecarboxaldehyde
CS-72545	C[C@@H](CO)N	(S)-(+)-2-amino-1-propanol
CS-13835336	C(CO)N	2-aminoethanol
CS-13854944	CN(C)CCO	Dimethylethanolamine
CS-13837537	C1COCCN1	Morpholine
CS-13836021	CNCCO	N-methylethanolamine
CS-13835861	CC(C)(CO)N	2-amino-2-methyl-1-propanol
CS-10309795	c1cn[nH]c1C=O	1H-pyrazole-3-carbaldehyde
CS-8373112	C1=CON=N1	Oxadiazole
CS-8009726	[O-]\\[N+](=N/C)CO	Hydroxymethyl-methylimino-oxido-ammonium
CS-4481813	C/C=N/O	Acetaldoxime
CS-311940	c1cn[nH]c1O	Pyrazol-5-ol
CS-283071	Cc1cc(n[nH]1)O	3-Methyl-5-pyrazolone
CS-161778	c1coc(=O)[nH]1	Oxazolone
CS-110510	C=C(C(=O)O)N	Dehydroalanine
CS-86862	c1c([nH]nc1O)N	3-Amino-5-pyrazolone
CS-66243	C(C(CO)O)N	1-aminoglycerol
CS-65711	CS(=O)(=O)N	Methanesulfonamide
CS-59559	Cc1cc(no1)N	3-amino-5-methyl-isoxazole
CS-55104	C(C(CN)O)N	1,3-diamino-2-propanol
CS-13867413	COC(=O)CC#N	Methyl cyanoacetate
CS-61591	C(C(CO)N)O	Serinol
CS-30663	C/N=[N+](\\C)/[O-]	Azoxymethane
CAFF	Cn1cnc2n(C)c(=O)n(C)c(=O)c12	Caffeine
Water	O	Water, TIP3, H charge only
WaterUCD	O	Water, TIP3, H charge and LJ
METHANAL	C=O	Formaldehyde
THEOBROMINE	Cn1cnc2c1c(=O)[nH]c(=O)n2C	Theobromine
BENZENE	c1ccccc1	Benzene
FURAN	c1ccoc1	Furan
PYRROLE	[nH]1cccc1	Pyrrole

Table S3 Additional chemicals generated via their SMILES code using ACPYPE and parameterised for use in the PMF Prediction toolkit, part 2
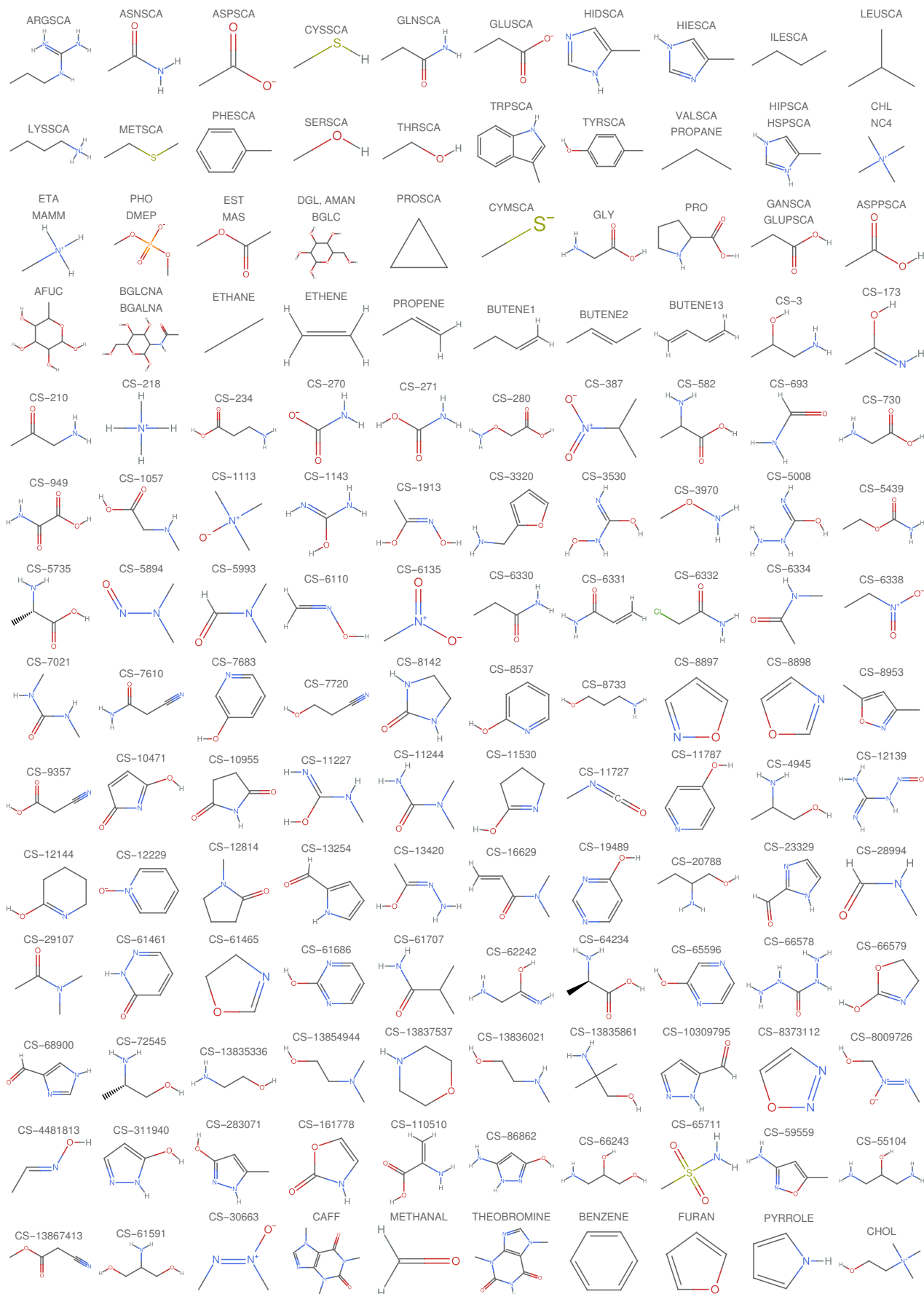
Figure S1 ID tags and structures for the full set of chemicals parameterised for the prediction of PMFs. Molecules sharing a SMILES code are grouped but may possess different structures or force field parameterisations.

Surface ID	Shape	Source	SSD Type (s)	Resolution [nm]	Notes
AuFCC100	Planar	0	0	0.0032	FCC Gold (100)
CdSeWurtzite2-10	Planar	0	0	0.0045	CdSe Wurtzite (2-10)
TiO2-ana-100	Planar	1	1	0.002	Hydrated anatase (100)
TiO2-ana-101	Planar	1	1	0.002	Hydrated anatase (101)
TiO2-rut-110	Planar	1	1	0.002	Hydrated rutile (110)
TiO2-rut-100	Planar	1	1	0.002	Hydrated rutile (110)
Fe2O3-001O	Planar	1	1	0.002	Hydrated iron oxide (001)
SiO2-Quartz	Planar	1	1	0.002	Quartz
SiO2-Amorphous	Planar	1	1	0.002	Amorphous silica
CNT15	Cylindrical	1	0	0.002	Carbon nanotube, 1.5 nm diameter
CNT15-COOH-30	Cylindrical	1	0	0.002	Carbon nanotube, 1.5 nm diameter, 30%wt COOH modified
CNT15-COOH-3	Cylindrical	1	0	0.002	Carbon nanotube, 1.5 nm diameter, 3%wt COOH modified
CNT15-COO–10	Cylindrical	1	0	0.002	Carbon nanotube, 1.5 nm diameter, 10%wt COO^- modified
CNT15-COO–3	Cylindrical	1	0	0.002	Carbon nanotube, 1.5 nm diameter, 3%wt COO^- modified
CNT15-NH2-14	Cylindrical	1	0	0.002	Carbon nanotube, 1.5 nm diameter, 14%wt NH_2 modified
CNT15-NH2-2	Cylindrical	1	0	0.002	Carbon nanotube, 1.5 nm diameter, 2%wt NH_2 modified
CNT15-NH3+-4	Cylindrical	1	0	0.002	Carbon nanotube, 1.5 nm diameter, 4%wt NH_3^+ modified
CNT15-NH3+-2	Cylindrical	1	0	0.002	Carbon nanotube, 1.5 nm diameter, 2%wt NH_3^+ modified
CNT15-OH-14	Cylindrical	1	0	0.002	Carbon nanotube, 1.5 nm diameter, 14%wt OH modified
CNT15-OH-4	Cylindrical	1	0	0.002	Carbon nanotube, 1.5 nm diameter, 4%wt OH modified
graphene	Planar	1	3	0.002	Single-layer graphene
bi-graphene	Planar	1	3	0.002	Double-layer graphene
tri-graphene	Planar	1	3	0.002	Triple-layer graphene
grapheneoxide	Planar	1	3	0.002	Graphene oxide, 30% oxidation
redgrapheneoxide	Planar	1	3	0.002	Reduced graphene oxide, 10% oxidation
C-amorph-2	Planar	1	1	0.002	Amorphous carbon
AuFCC100UCD	Planar	3	2	0.00051	FCC Gold (100)
AuFCC110UCD	Planar	2	2	0.00034	FCC Gold (110)
AuFCC111UCD	Planar	2	2	0.00029	FCC Gold (111)
Ag100	Planar	3	2	0.0018	FCC Silver (100)
Ag110	Planar	2	2	0.0018	FCC Silver (101)
Ag111	Planar	2	2	0.0018	FCC Silver (111)
Fe001	Planar	3	2	0.0018	FCC Iron (001/100)
Fe110	Planar	2	2	0.0019	FCC Iron (110)
Fe111	Planar	2	2	0.0019	FCC Iron (111)
Cu001	Planar	3	2	0.00030	FCC Copper (001/100)
Cu110	Planar	2	2	0.00021	FCC Copper (110)
Cu111	Planar	2	2	0.00025	FCC Copper (111)
C-amorph-3	Planar	1	1	0.002	Amorphous carbon
AuFCC100-Ablate0	Planar	1	0	0.002	FCC Gold (100), 0% ablation
AuFCC100-Ablate5	Planar	1	0	0.002	FCC Gold (100), 5% ablation
AuFCC100-Ablate25	Planar	1	0	0.002	FCC Gold (100), 25% ablation
AuFCC100-Ablate50	Planar	1	0	0.002	FCC Gold (100), 50% ablation
AuFCC100-Ablate75	Planar	1	0	0.002	FCC Gold (100), 75% ablation
AuFCC100-Ablate100	Planar	1	0	0.002	FCC Gold (100), 100% ablation
GoldBrush	Planar	1	0	0.002	FCC Gold (100), PE brush, 50% coverage
CaO001	Planar	1	0	0.002	FCC Calcium oxide (001)
Pt001	Planar	1	0	0.002	FCC Platinum(001/100)
TricalciumSilicate001	Planar	1	0	0.002	Tricalcium silicate (001)
Au-001-PE	Planar	1	0	0.002	Au (100), PE brush, 100% coverage
Au-001-PEG	Planar	1	0	0.002	Au (100), PEG brush, 100% coverage
Al2O3-001	Planar	1	0	0.002	Aluminium oxide (001)
Cr2O3-001	Planar	1	0	0.002	Chromium oxide (001)
Ce-001	Planar	1	0	0.002	FCC Cerium (001)
Hydroxyapatite-001	Planar	1	0	0.002	Hydroxyapatite (001)

Table S4 A summary of the surfaces considered in this work, listing the internal ID, shape, source (0: SU (no ions), 1: SU (ions), 2: UCD-1 (110 and 111 surfaces), 3: UCD-2 (100 surfaces)) and the zero-indexed categorical variable defining the convention for the SSD (0: nominal surface - COM distance, 1: minimum surface atom - COM distance, 2: slab-width-adjusted COM - COM distance, 3: surface atom COM - COM distance), with e.g. $s = 0$ corresponding to SSD class 1 in the main text. The first set indicates development surfaces, the second testing surfaces, and the final are additional surfaces without known PMFs.

Model ID	Class	E_{ads} Correlation	E_{ads} R^2	$\langle KL \rangle$
Cluster-A-1	TMTC	1.0	0.98	0.17
Cluster-A-1	TMVC	0.9	0.78	0.37
Cluster-A-1	VMTC	0.94	0.86	0.57
Cluster-A-1	VMVC	0.83	0.62	0.69
Cluster-A-2	TMTC	1.0	0.99	0.16
Cluster-A-2	TMVC	0.89	0.79	0.33
Cluster-A-2	VMTC	0.79	0.44	0.71
Cluster-A-2	VMVC	0.83	0.67	0.58
Cluster-A-3	TMTC	1.0	0.99	0.16
Cluster-A-3	TMVC	0.73	0.5	0.36
Cluster-A-3	VMTC	0.7	0.36	0.37
Cluster-A-3	VMVC	0.67	0.01	0.33
Cluster-A-4	TMTC	1.0	0.98	0.16
Cluster-A-4	TMVC	0.82	0.66	0.37
Cluster-A-4	VMTC	0.88	0.53	0.55
Cluster-A-4	VMVC	0.65	0.34	0.87
Cluster-A-5	TMTC	1.0	0.98	0.15
Cluster-A-5	TMVC	0.9	0.72	0.43
Cluster-A-5	VMTC	0.93	0.83	0.38
Cluster-A-5	VMVC	0.88	0.76	0.36
Simple-B-1	T	1.0	0.99	0.14
Simple-B-1	V	0.91	0.82	0.39
Simple-B-2	T	1.0	0.99	0.15
Simple-B-2	V	0.91	0.83	0.32
Simple-B-3	T	1.0	0.99	0.14
Simple-B-3	V	0.9	0.82	0.41
Simple-B-4	T	1.0	0.98	0.14
Simple-B-4	V	0.89	0.78	0.4
Simple-B-5	T	1.0	0.99	0.15
Simple-B-5	V	0.91	0.83	0.39
Simple-B-6	T	1.0	0.99	0.13
Simple-B-6	V	0.94	0.87	0.33
Simple-B-7	T	1.0	0.99	0.14
Simple-B-7	V	0.93	0.85	0.33
Simple-B-8	T	1.0	0.99	0.14
Simple-B-8	V	0.87	0.74	0.35
Simple-B-9	T	1.0	0.99	0.14
Simple-B-9	V	0.92	0.84	0.31
Simple-B-10	T	1.0	0.99	0.2
Simple-B-10	V	0.93	0.86	0.43

Table S5 Summary statistics for the accuracy of binding energies (correlation and R^2) and mean Kullback-Leibler divergences for PMFs extracted from the predicted potentials of mean force compared to the input potentials for the training and validation groups for the model versions summarised by a split type (clustering then random or simple random), the use of all data (A) or bootstrap data (B) and an ID number. For simple random splitting, the class column indicates results for PMFs included in the training set for that particular bootstrap dataset and training-validation split, with the validation set containing all out-of-bag PMFs. For cluster splitting, TMTC indicates the training material, training chemical set and so on, with all PMFs featured in one of these four classes.

Material	Chemical	E(MD) [kJ·mol^{-1}]	E(Simple) [kJ·mol^{-1}]	E(Cluster) [kJ·mol^{-1}]
Ag100	BGALNA-JS	−49.	−13.7±2.7	−13.5±2.1
Ag100	CHOL-JS	−10.6	−7.9±2.9	−3.3±0.8
Ag110	BGALNA-JS	−64.4	−32.±4.	−31.2±2.9
Ag110	CHOL-JS	−21.2	−29.±6.	−27.±5.
Ag111	BGALNA-JS	−68.8	−46.±6.	−43.±5.
Ag111	CHOL-JS	−15.8	−34.±11.	−38.±14.
AuFCC100UCD	BGALNA-JS	−15.5	−14.±4.	−12.0±2.0
AuFCC100UCD	CHOL-JS	−7.8	−7.±4.	−5.3±0.7
AuFCC110UCD	BGALNA-JS	−51.	−36.±4.	−33.±6.
AuFCC110UCD	CHOL-JS	−20.2	−30.±6.	−32.±13.
AuFCC111UCD	BGALNA-JS	−45.1	−45.±9.	−45.6±2.3
AuFCC111UCD	CHOL-JS	−18.5	−32.±8.	−33.±12.
Cu001	BGALNA-JS	−67.2	−36.±6.	−31.±4.
Cu110	BGALNA-JS	−51.3	−33.±9.	−27.5±3.2
Cu111	BGALNA-JS	−65.3	−46.±8.	−47.6±2.2

Table S6 Adsorption energies of two molecules not used in the training set to FCC gold, silver and copper for (100), (110) and (111) surfaces, comparing the values found from metadynamics simulations to those predicted by the two ensemble models. The provided errors are obtained from the standard deviation of the adsorption energies predicted by each individual ensemble member.

Material	Chemical	E(MD) [kJ·mol^{-1}]	E(Simple) [kJ·mol^{-1}]	E(Cluster) [kJ·mol^{-1}]
C-amorph-3	ALASCA-AC	−2.2	−0.3±0.5	0.60±0.19
C-amorph-3	ARGSCA-AC	−16.5	−16.9±1.1	−11.7±2.3
C-amorph-3	ASNSCA-AC	−8.5	−6.9±1.5	−5.2±2.0
C-amorph-3	ASPSCA-AC	−4.1	−2.3±1.2	−1.8±1.0
C-amorph-3	CHL-AC	−3.1	−2.1±0.9	−1.47±0.33
C-amorph-3	CYMSCA-AC	−2.2	−1.5±1.2	−0.9±0.8
C-amorph-3	CYSSCA-AC	−5.1	−2.0±0.8	−0.61±0.33
C-amorph-3	DGL-AC	−19.	−17.1±2.6	−14.±4.
C-amorph-3	EST-AC	−13.5	−12.4±1.4	−9.2±1.6
C-amorph-3	ETA-AC	−0.7	0.0±0.5	0.35±0.33
C-amorph-3	GANSCA-AC	−12.8	−11.9±1.6	−8.7±1.6
C-amorph-3	GLNSCA-AC	−11.3	−9.7±1.7	−7.2±1.3
C-amorph-3	GLUSCA-AC	−6.5	−7.1±3.0	−5.6±2.1
C-amorph-3	GLY-AC	−9.8	−8.4±1.9	−7.4±2.5
C-amorph-3	HIDSCA-AC	−14.8	−14.9±1.2	−11.5±2.5
C-amorph-3	HIESCA-AC	−14.9	−15.9±0.8	−12.9±1.9
C-amorph-3	HIPSCA-AC	−12.7	−15.3±1.7	−11.±4.
C-amorph-3	ILESCA-AC	−11.9	−8.0±1.9	−5.9±1.6
C-amorph-3	LEUSCA-AC	−11.8	−7.3±2.7	−5.6±0.9
C-amorph-3	LYSSCA-AC	−9.	−5.3±1.7	−3.8±1.4
C-amorph-3	METSCA-AC	−12.4	−9.5±2.1	−7.5±2.4
C-amorph-3	PHESCA-AC	−16.5	−15.7±2.4	−10.7±0.5
C-amorph-3	PHO-AC	−8.	−6.5±2.7	−6.8±2.3
C-amorph-3	PRO-AC	−18.4	−15.6±2.4	−11.3±2.7
C-amorph-3	SERSCA-AC	−3.9	−1.2±0.7	0.0±0.4
C-amorph-3	THRSCA-AC	−6.9	−3.1±0.9	−2.6±1.0
C-amorph-3	TRPSCA-AC	−21.1	−25.3±1.9	−19.0±2.5
C-amorph-3	TYRSCA-AC	−20.2	−22.6±2.8	−15.5±3.0
C-amorph-3	VALSCA-AC	−8.8	−5.1±1.6	−3.0±1.3

Table S7 Adsorption energies of molecules used for the development of the model to an amorphous carbon surface not included in the training set, comparing the values found from metadynamics simulations to those predicted by the two ensemble models. The provided errors are obtained from the standard deviation of the adsorption energies predicted by each individual ensemble member.

Material	Chemical	E(MD) [kJ·mol^{-1}]	E(Simple) [kJ·mol^{-1}]	E(Cluster) [kJ·mol^{-1}]
Cu111	AFUC-JS	−43.1	−37. ± 10.	−39. ± 5.
Cu111	ALASCA-JS	1.7	−2.3 ± 2.6	0.4 ± 0.6
Cu111	AMAN-JS	−55.7	−41. ± 10.	−55. ± 5.
Cu111	ARGSCA-JS	−72.	−52. ± 12.	−53. ± 18.
Cu111	ASNSCA-JS	−20.1	−19.0 ± 2.0	−18.4 ± 1.2
Cu111	ASPPSCA-JS	−23.6	−18.4 ± 1.7	−17.6 ± 1.4
Cu111	ASPSCA-JS	−12.2	−10.4 ± 1.7	−9.5 ± 0.4
Cu111	BGALNA-JS	−65.3	−46. ± 8.	−47.6 ± 2.2
Cu111	BGLC-JS	−55.9	−48. ± 9.	−56.3 ± 3.4
Cu111	BGLCNA-JS	−76.	−45. ± 6.	−44. ± 5.
Cu111	CYMSCA-JS	−22.3	−18.0 ± 1.6	−17.2 ± 0.9
Cu111	CYSSCA-JS	−14.8	−13.8 ± 1.3	−12.0 ± 0.9
Cu111	DMEP-JS	−57.5	−39. ± 6.	−45.4 ± 2.9
Cu111	GLNSCA-JS	−25.7	−21.5 ± 1.9	−21.6 ± 1.1
Cu111	GLUPSCA-JS	−29.9	−20.8 ± 2.2	−20.2 ± 2.8
Cu111	GLUSCA-JS	−14.8	−14. ± 4.	−11.1 ± 3.0
Cu111	HIDSCA-JS	−31.8	−27. ± 4.	−26.2 ± 1.3
Cu111	HIESCA-JS	−33.6	−26.3 ± 2.8	−24.3 ± 1.5
Cu111	HIPSCA-JS	−42.	−30. ± 4.	−29. ± 5.
Cu111	ILESCA-JS	−10.1	−12.0 ± 2.5	−12.4 ± 3.4
Cu111	LEUSCA-JS	−4.	−9.8 ± 3.1	−8.1 ± 1.4
Cu111	LYSSCA-JS	−12.3	−18. ± 7.	−15.1 ± 3.1
Cu111	MAMM-JS	−3.8	−2.8 ± 1.3	−2.2 ± 2.1
Cu111	MAS-JS	−23.6	−20.3 ± 2.5	−18.9 ± 2.8
Cu111	METSCA-JS	−27.8	−23.9 ± 2.3	−23.0 ± 0.9
Cu111	NC4-JS	−14.3	−15.0 ± 2.0	−13.0 ± 1.4
Cu111	PHESCA-JS	−32.7	−28. ± 4.	−27.8 ± 1.0
Cu111	PROSCA-JS	−4.	−8.6 ± 1.5	−6.8 ± 1.0
Cu111	SERSCA-JS	−3.	−5.0 ± 3.1	−3.4 ± 2.4
Cu111	THRSCA-JS	−5.9	−9. ± 4.	−8.6 ± 1.7
Cu111	TRPSCA-JS	−75.	−54. ± 11.	−60.6 ± 3.1
Cu111	TYRSCA-JS	−51.1	−35. ± 8.	−38. ± 9.
Cu111	VALSCA-JS	−1.6	−8.3 ± 2.7	−6.2 ± 0.8

Table S8 Adsorption energies of molecules used for the development of the model to a copper (111) surface not included in the training set, comparing the values found from metadynamics simulations to those predicted by the two ensemble models. The provided errors are obtained from the standard deviation of the adsorption energies predicted by each individual ensemble member.
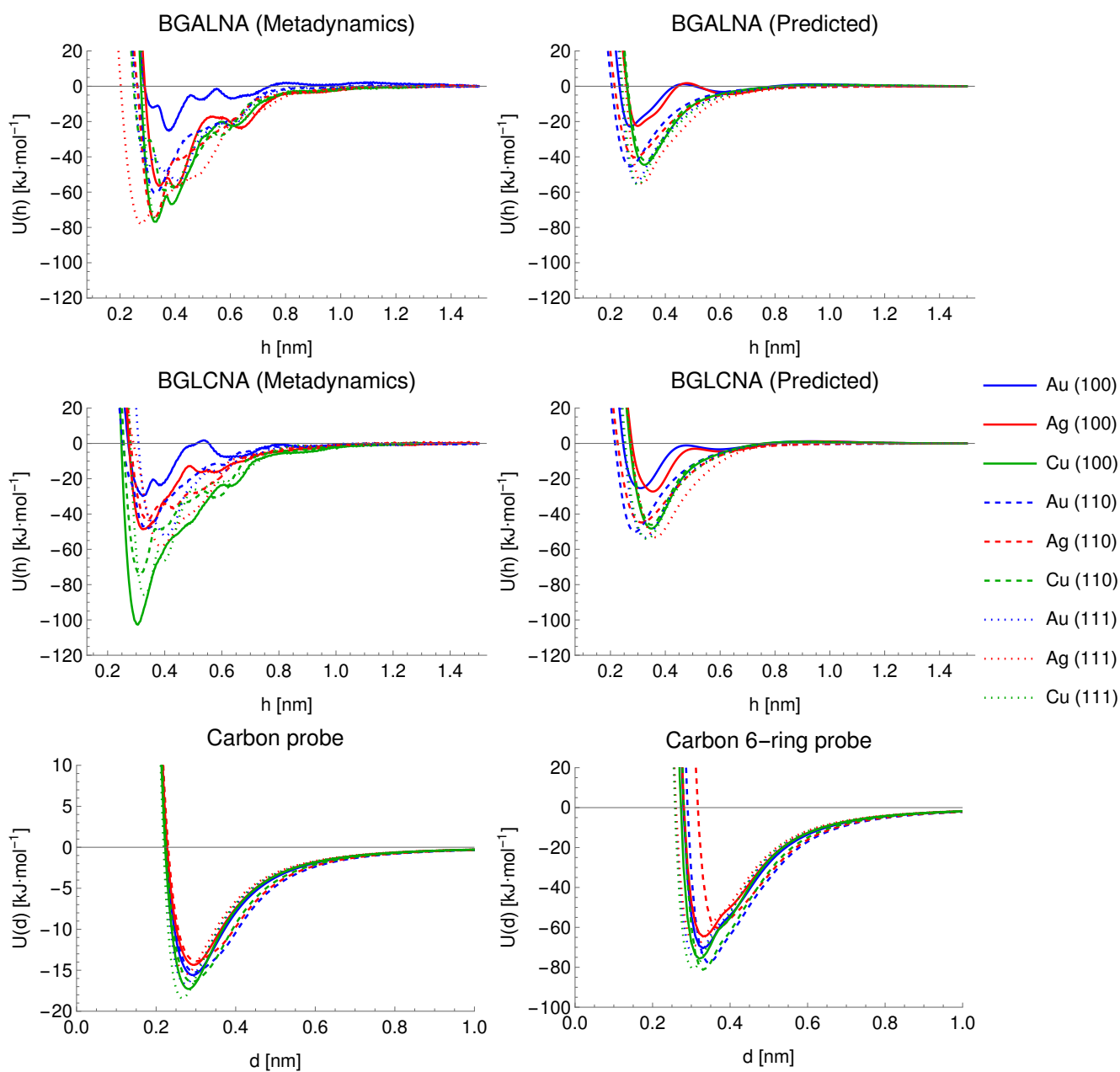
Figure S2 Top: Potentials of mean force for the testing sugar molecule BGALNA on nine FCC metal surfaces obtained via metadyamics simulations (left) and using the approach detailed in this work (right). Prediction parameters are matched to those used for the metadynamics simulations. Middle: As top, but for the sugar molecule BGLCNA used in the training set. Bottom: Input potentials for these FCC metal surfaces showing the single atom carbon probe (left) and the six-membered ring probe (right). In all cases, the gold (blue) and silver (red) surfaces are used in training while copper (green) was reserved for testing.

Figure S3 Generated potentials of mean force for the Au FCC (100) surface, showing the canonical form (blue), match to original parameters (red, dashed) and input PMF where available (green).
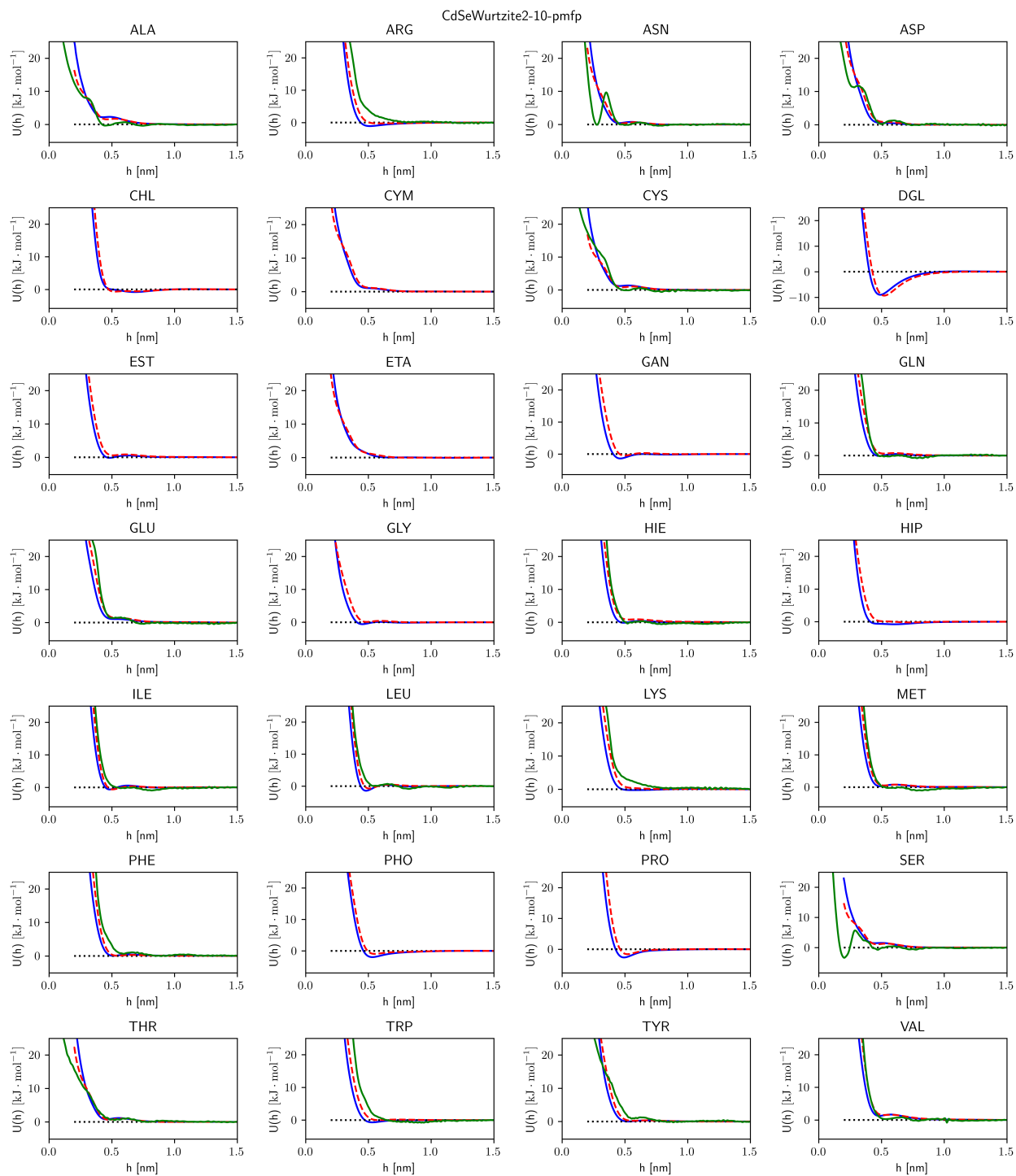
Figure S4 Generated potentials of mean force for the CdSe wurtzite (2-10) surface, showing the canonical form (blue), match to original parameters (red, dashed) and input PMF where available (green).
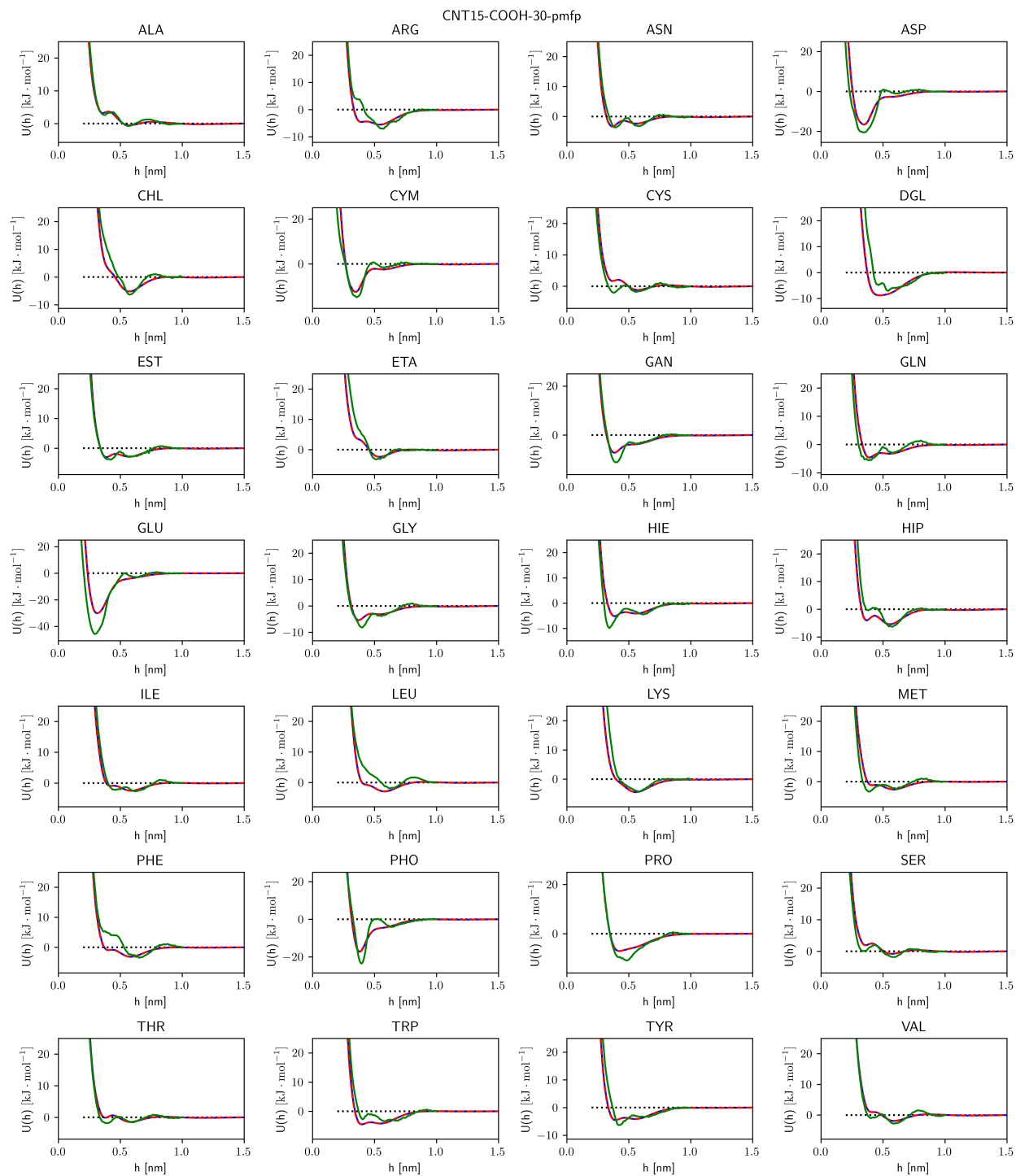
Figure S5 Generated potentials of mean force for the 30% by weight COOH modified carbon nanotube surface, showing the canonical form (blue), match to original parameters (red, dashed) and input PMF where available (green).
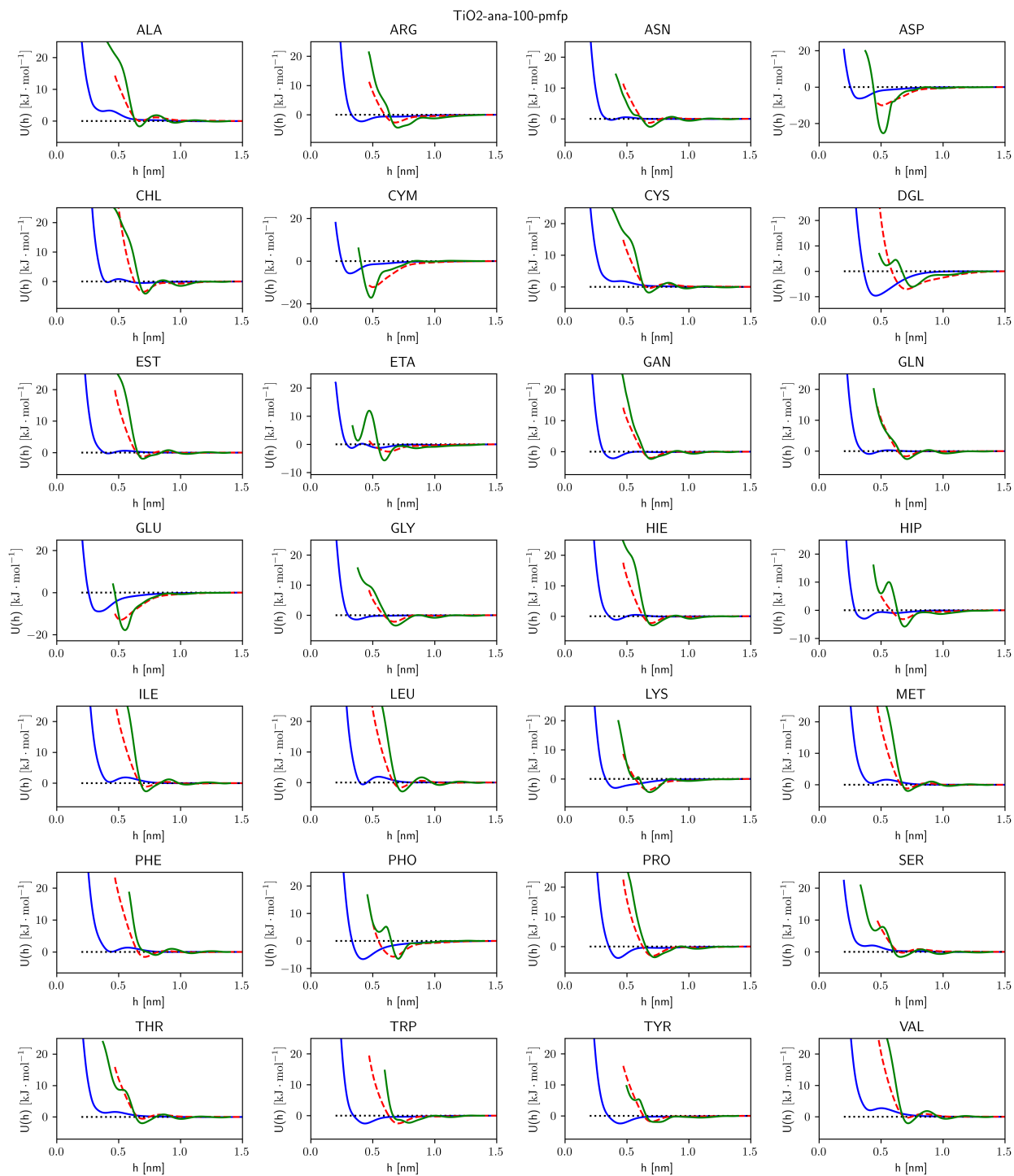
Figure S6 Generated potentials of mean force for the titanium dioxide anatase (100) surface, showing the canonical form (blue), match to original parameters (red, dashed) and input PMF where available (green).