Supplementary Information (SI) for:

Advancing Density Functional Tight-Binding Method for Large Organic Molecules through Equivariant Neural Networks

Leonardo Medrano Sandonas,[*1] Mirela Puleva,[2,3] Zekiye Erarslan,[1] Ricardo Parra Payano,[4] Martin Stöhr,[5,6] Gianaurelio Cuniberti,[1] and Alexandre Tkatchenko[*2,3]

[1] *Institute for Materials Science and Max Bergmann Center of Biomaterials, TUD Dresden University of Technology, 01062 Dresden, Germany*
[2] *Department of Physics and Materials Science, University of Luxembourg, L-1511 Luxembourg City, Luxembourg*
[3] *Institute for Advanced Studies, University of Luxembourg, Campus Belval, L-4365 Esch-sur-Alzette, Luxembourg*
[4] *Universidad Nacional de Ingeniería, Av. Túpac Amaru 210, Rímac, Lima 15333, Peru.*
[5] *Department of Chemistry and The PULSE Institute, Stanford University, Stanford CA-94305, USA.*
[6] *SLAC National Accelerator Laboratory, Menlo Park CA-94025, USA.*

[*] Corresponding authors: Leonardo Medrano Sandonas (`leonardo.medrano@tu-dresden.de`), Alexandre Tkatchenko (`alexandre.tkatchenko@uni.lu`)

1 Optimization of equivariant potentials

Table S1 Analysis of the performance of Δ_{TB} potentials trained with SpookyNet (SP), Allegro (AG), and MACE (MC) on small single molecules from the QM7-X dataset (label '1' after the equivariant NN). We compute the mean absolute error (MAE) for the prediction of ΔE_{TB} and $\Delta\mathbf{F}$TB for molecular conformations in the QM7-X and DES15K datasets. Additionally, we report the MAE, root-mean-squared error (RMSE), and mean absolute relative error (MARE) for the prediction of interaction energies of molecular dimers from the S66X8 dataset. We also include results for ΔTB potentials trained using the short-range module (*e.g.*, ZBL repulsion interaction) implemented in SpookyNet. Furthermore, we optimized the MC1 model with respect to the cutoff radius (r_c), identifying $r_c = 4.0$ Å as the value that parameterizes the best-performing EquiDTB model with MACE. Errors for energies and atomic forces are given in kcal/mol and kcal/mol·Å, respectively.

Δ_{TB}	Hyperparameter			QM7-X		DES15K		S66x8		
potential	N_{int}	l_{max}	r_c	ΔE_{TB}	$\Delta\mathbf{F}_{\mathrm{TB}}$	ΔE_{TB}	$\Delta\mathbf{F}_{\mathrm{TB}}$	MAE	RMSE	MARE
SP1	2	2	5.0	0.42	1.24	15.11	6.78	2.16	3.54	111.1
SP1+ZBL	2	2	5.0	0.37	1.09	15.85	6.41	1.35	2.34	120.7
AG1	2	2	5.0	0.21	0.24	14.97	5.31	1.92	2.95	61.0
	2	2	3.0	0.50	0.48	14.49	5.31	0.81	1.06	35.4
MC1	2	2	4.0	0.37	0.38	14.69	5.15	0.97	1.16	33.9
	2	2	5.0	0.38	0.39	14.67	5.15	1.72	1.98	62.7

Table S2 Analysis of the performance of Δ_{TB} potentials trained with SpookyNet (SP), Allegro (AG), and MACE (MC) on small single molecules and molecular dimers from the QM7-X and DES15K datasets (label '2' after the equivariant NN). We compute the mean absolute error (MAE) for the prediction of ΔE_{TB} and $\Delta\mathbf{F}$TB for molecular conformations in the QM7-X and DES15K datasets. Additionally, we report the MAE, root-mean-squared error (RMSE), and mean absolute relative error (MARE) for the prediction of interaction energies of molecular dimers from the S66X8 dataset. We also include results for ΔTB potentials trained using the short-range module (*e.g.*, ZBL repulsion interaction) implemented in SpookyNet. Furthermore, we optimized the MC2 model with respect to the cutoff radius (r_c), identifying $r_c = 4.0$ Å as the value that parameterizes the best-performing EquiDTB model with MACE. Errors for energies and atomic forces are given in kcal/mol and kcal/mol·Å, respectively.

Δ_{TB}	Hyperparameter			QM7-X		DES15K		S66x8		
potential	N_{int}	l_{max}	r_c	ΔE_{TB}	$\Delta\mathbf{F}_{\mathrm{TB}}$	ΔE_{TB}	$\Delta\mathbf{F}_{\mathrm{TB}}$	MAE	RMSE	MARE
SP2	2	2	5.0	0.37	1.17	0.30	6.12	0.91	1.14	44.1
SP2+ZBL	2	2	5.0	0.36	1.09	0.30	5.89	1.02	1.24	48.5
AG2	2	2	5.0	0.33	0.34	1.81	0.71	1.19	1.49	46.0
	2	2	3.0	0.54	0.52	4.19	2.42	1.11	1.78	38.1
MC2	2	2	4.0	0.44	0.44	3.65	2.18	1.47	1.99	41.2
	2	2	5.0	0.43	0.42	3.42	2.00	1.67	2.06	57.1

Table S3 Analysis of the performance of the reference machine learning potential (rMLP) trained with MACE on molecular conformations from only QM7-X dataset and both datasets. We compute the mean absolute error (MAE) for the prediction of total energies E and atomic forces $\mathbf{F}$ at PBE0+MBD level of molecules in the QM7-X and DES15K datasets. Additionally, we report the MAE, root-mean-squared error (RMSE), and mean absolute relative error (MARE) for the prediction of interaction energies of molecular dimers from the S66X8 dataset. We also optimized the rMLP potentials with respect to the cutoff radius (r_c), identifying $r_c = 5.0$ Å as the value that parameterizes the best-performing rMLP potential with MACE, which was trained on both datasets. Errors for energies and atomic forces are given in kcal/mol and kcal/mol·Å, respectively.

Training	Hyperparameter			QM7-X		DES15K		S66x8		
dataset	N_{int}	l_{max}	r_c	E	$\mathbf{F}$	E	$\mathbf{F}$	MAE	RMSE	MARE
QM7-X	2	2	3.0	0.59	0.63	20.98	2.25	2.62	3.19	82.2
	2	2	4.0	0.49	0.53	20.31	2.08	2.49	3.01	97.5
	2	2	5.0	0.47	0.54	19.29	2.11	1.83	2.36	85.7
	2	2	6.0	0.49	0.55	19.09	2.20	2.10	2.47	90.2
QM7-X and DES15K	2	2	3.0	0.62	0.67	5.94	1.78	2.52	3.00	80.5
	2	2	4.0	0.50	0.57	5.10	1.60	2.12	2.66	75.9
	2	2	5.0	0.48	0.57	4.87	1.57	1.19	1.51	48.0
	2	2	6.0	0.52	0.59	5.28	1.64	1.52	1.88	65.0

2 Additional results for non-covalent systems

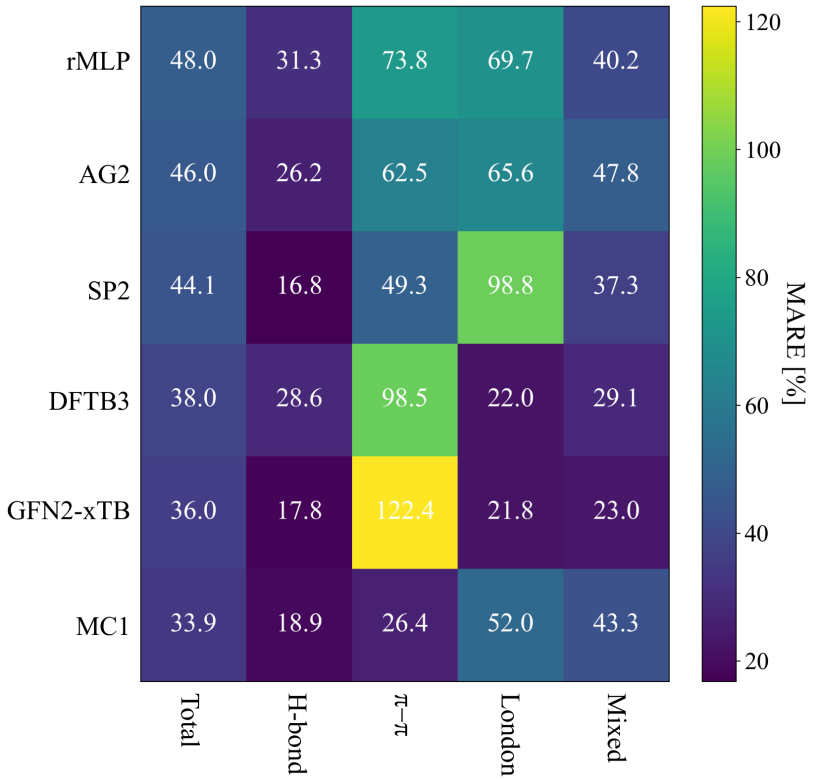

Fig. S1 Benchmarking Δ_{TB} potentials for calculating interaction energies (E_{int}) in small molecular dimers from the S66x8 dataset. Heatmap plot of the mean absolute relative errors (MARE) for each studied model, split according to the predominant non-covalent interaction in the molecular dimer. Reference values for E_{int} were calculated using PBE0+MBD. All calculations include a many-body dispersion treatment, except for xTB, which considers D4 correction.

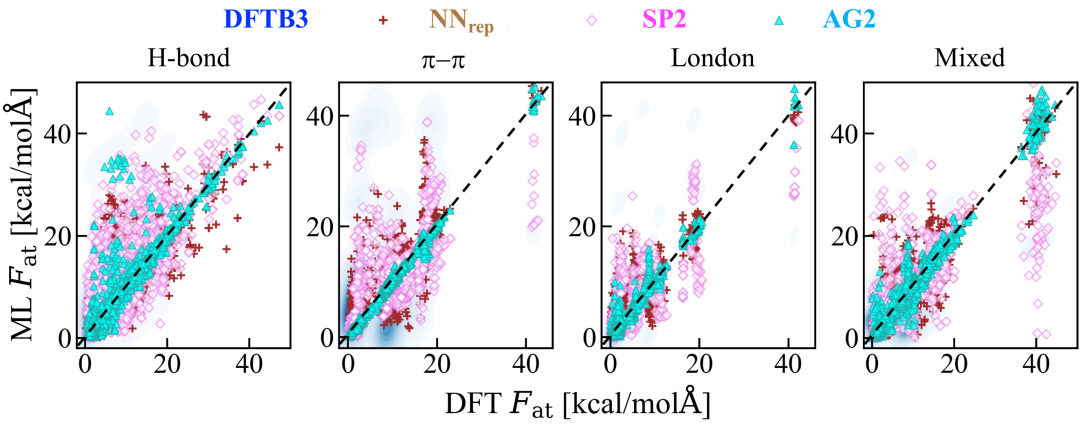

Fig. S2 Benchmarking Δ_{TB} potentials for calculating atomic forces ($\mathbf{F}_{\mathrm{at}}$) in small molecular dimers from the S66x8 dataset. Correlation plots between DFT and ML F_{at} values for each dimer group computed by using the $\mathrm{NN}_{\mathrm{rep}}$, SP2 and AG2 models. For comparison, we also include the corresponding values obtained with DFTB3. Reference values for $\mathbf{F}_{\mathrm{at}}$ were calculated using PBE0+MBD. All calculations include a many-body dispersion treatment.

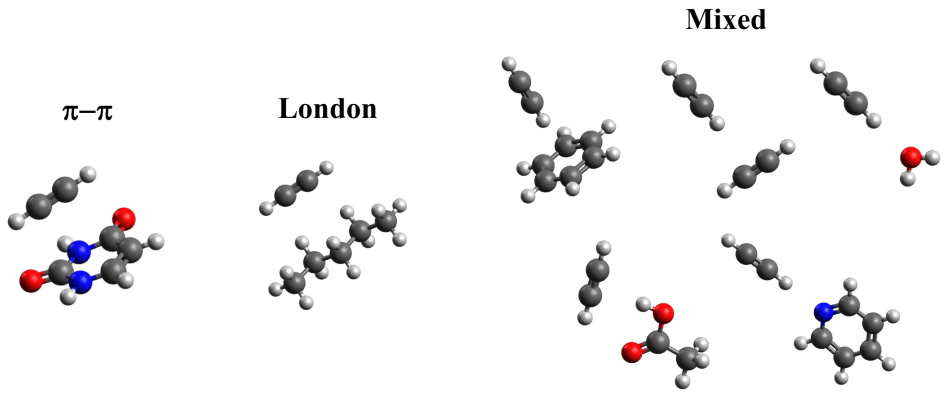

Fig. S3 Ethyne-based molecular dimers, for which the standard tight-binding methods (DFTB3 and GFN2-xTB) failed to accurately compute their atomic forces, $\mathbf{F}_{\mathrm{at}}$. Reference values for $\mathbf{F}_{\mathrm{at}}$ were calculated using PBE0+MBD.

Table S4 Performance of Δ_{TB} potentials in predicting the interaction energies E_{int} and atomic forces F_{at} for equilibrium and non-equilibrium small molecular dimers in S66x8 dataset. We show the mean absolute errors (MAE) for the $\mathrm{NN_{rep}}$, SP2, AG2, and EquiDTB (*i.e.*, MC1) models. For comparison, the error values for widely used TB methods (DFTB3 and GFN2-xTB) and the reference ML potential (rMLP) are also presented. All calculations consider a many-body dispersion treatment, except for xTB, which considers D4 correction. The error values for energies and forces are given in kcal/mol and kcal/mol·Å, respectively.

Model	E_{int}			F_{at}
	Compressed	Elongated	Total	Total
DFTB3	1.39	0.83	1.04	4.52
GFN2-xTB	1.23	0.64	0.86	5.20
$\mathrm{NN_{rep}}$	3.45	2.94	3.13	1.57
SP2	0.99	0.86	0.91	1.68
AG2	1.45	1.03	1.19	0.55
EquiDTB	1.23	0.81	0.97	0.52
rMLP	1.25	1.16	1.19	0.57

3 Distribution of energy and force corrections

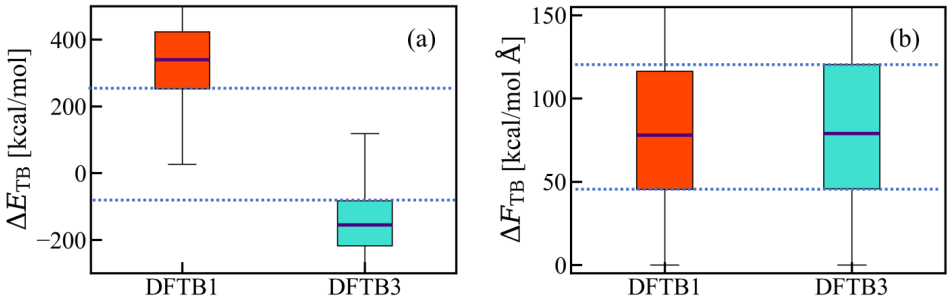

Fig. S4 Box plots for the target properties (a) ΔE_{TB} and (b) ΔF_{TB} obtained using the DFTB1 and DFTB3 methods.

4 Additional results for calculation of the minimum energy path

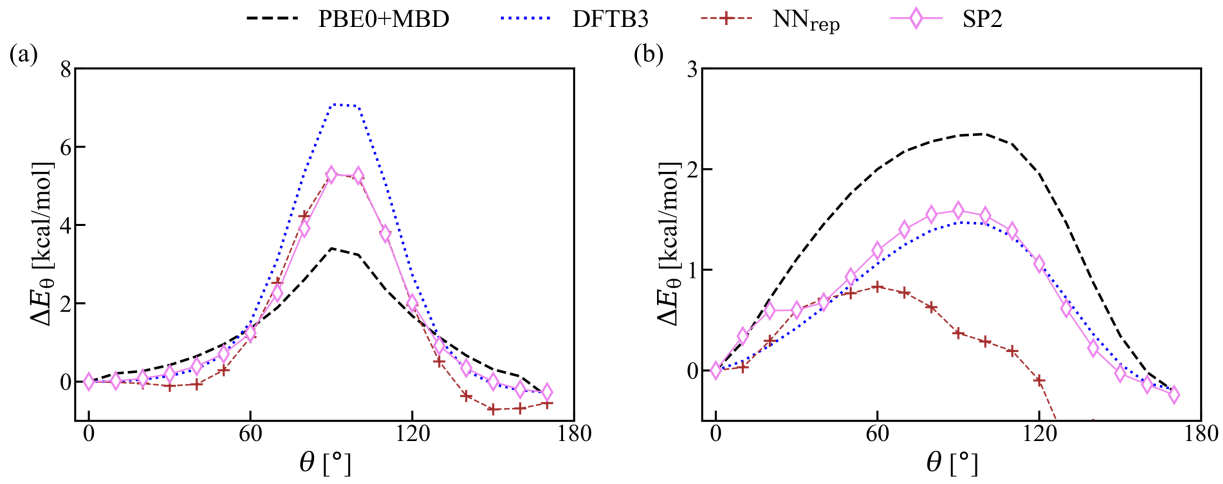

Fig. S5 Assessing predictions of non-equilibrium properties of flexible molecules. Minimum energy path for the rotation of the dihedral connecting the aromatic ring and the linear-type structure in (a) paracetamol and (b) tyrosine. The rotational profiles were computed performing Nudged Elastic Band (NEB) calculations. We present the results obtained by the DFTB3 method, NN_{rep} model, and SP2 model. Reference values for energies were calculated using PBE0+MBD. All calculations include a many-body dispersion treatment.

5 Energetic and structural properties along MD trajectories

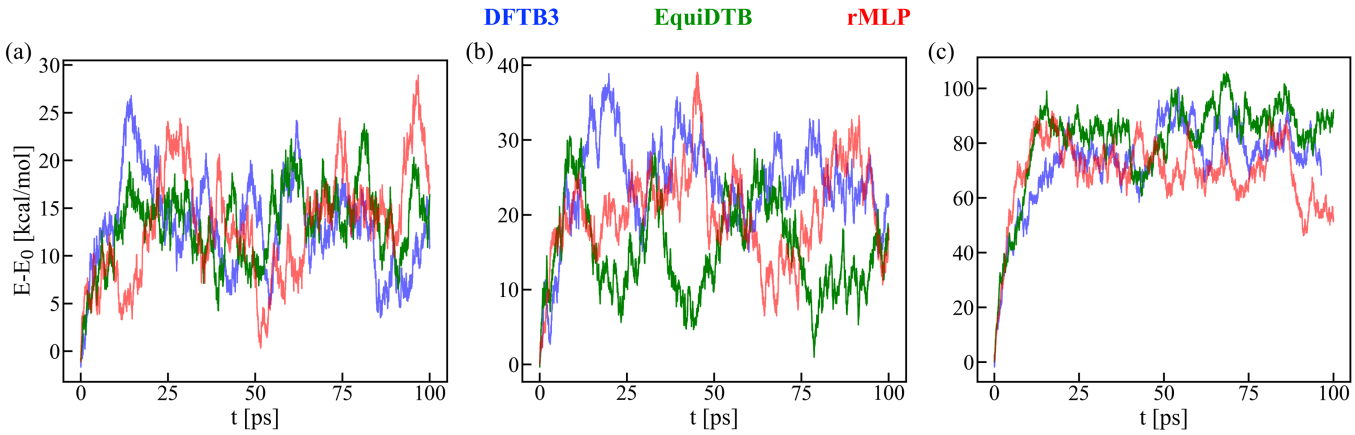

Fig. S6 Variation of the total energy of (a) alanine dipeptide, (b) zaprinast, and (c) ligand 2Q5k as a function of simulation time. Molecular dynamics simulations were performed at 300 K for 100 ps. Results are shown for the DFTB3 method, the EquiDTB model, and the rMLP potential. All calculations include a many-body dispersion treatment.

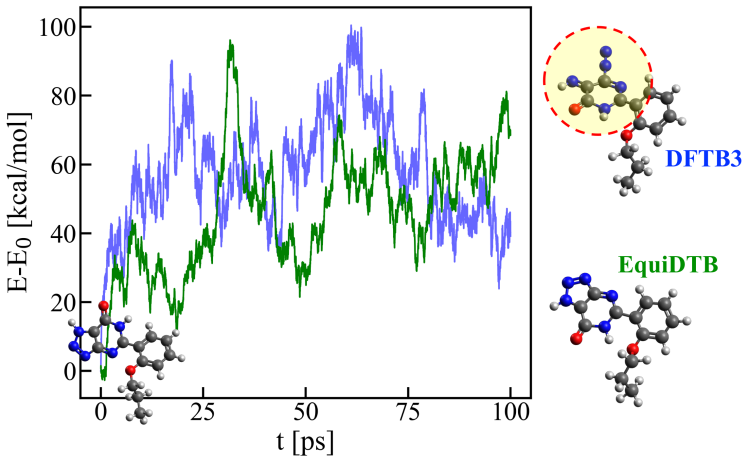

Fig. S7 Variation of the total energy of zaprinast as a function of simulation time at 600 K. Results are shown for the DFTB3 method and the EquiDTB model. Atomistic representations of the molecular structure of zaprinast at 80 ps of simulation are also provided to highlight bond-breaking in the triazole ring observed with the DFTB3 method.

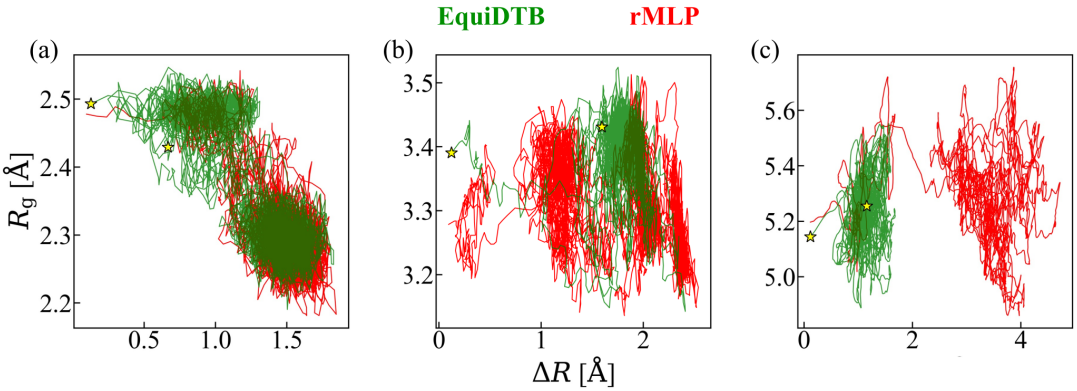

Fig. S8 To evaluate the structural evolution, we present the two-dimensional space defined by the root-mean-squared deviation with respect to the optimized geometry, ΔR, and the radius of gyration, R_g. The results are shown for (a) alanine dipeptide, (b) zaprinast, and (c) ligand 2Q5K, obtained using the EquiDTB model and the rMLP potential.

6 Computing vibrational modes of α-amino acids

Table S5 List of α-amino acids considered for the analysis of vibrational modes. The mean absolute error (MAE_ω) and the maximum deviation (ω_{max}) in frequency prediction per molecule are shown for the DFTB3 method, NN_{rep} model, and the SP2 model.

α-amino	Reduced	DFTB3		NN_{rep}		SP2	
acid	name	MAE_ω	ω_{max}	MAE_ω	ω_{max}	MAE_ω	ω_{max}
Alanine	Ala	55.5	294.0	6.09	21.7	21.7	189.3
Arginine	Arg	50.4	302.6	13.3	46.1	18.8	159.3
Asparagine	Asn	65.1	358.9	12.1	32.7	17.7	133.6
Aspartic acid	Asp	52.9	299.7	9.2	56.6	20.0	144.1
Glutamic acid	Glu	50.9	305.0	10.5	52.7	19.0	206.3
Glutamine	Gln	58.8	310.7	17.0	45.4	15.1	98.6
Glycine	Gly	57.4	290.6	7.1	32.0	26.1	215.0
Histidine	His	49.4	311.2	24.9	77.8	21.7	183.7
Isoleucine	Ile	47.1	292.5	7.5	26.8	13.3	96.1
Leucine	Leu	47.4	280.8	6.8	21.3	16.9	120.2
Lysine	Lys	46.5	285.4	5.6	27.0	8.8	66.5
Phenylalanine	Phe	50.5	286.0	15.3	51.7	21.5	160.3
Proline	Pro	43.9	322.0	10.3	34.0	20.2	199.4
Serine	Ser	62.1	283.1	9.3	47.6	18.3	93.9
Threonine	Thr	58.4	289.4	9.8	49.7	18.1	94.8
Tryptophan	Trp	48.0	333.1	19.2	90.9	16.0	152.8
Tyrosine	Tyr	47.8	285.5	16.0	44.9	16.1	128.1
Valine	Val	48.3	307.9	13.7	66.5	19.5	172.4

7 Determining the energetic ranking of drug-like molecules

Table S6 Performance of Δ_{TB} potentials in predicting the energetic ranking (R_{K}) for equilibrium and non-equilibrium conformations of large drug-like molecules extracted from Aquamarine dataset. We show the average of the mean absolute error for the change in energetic ranking place ΔR_{K} ($\langle \mathrm{MAE}\ \Delta R_{\mathrm{K}} \rangle$), the average of standard deviation of ΔR_{K} (σ), and the average of the maximum ΔR_{K} ($\langle (\Delta R_{\mathrm{K}})_{max} \rangle$).In this analysis, we consider ΔR_{K} values obtained using the widely used tight-binding methods (DFTB3 and GFN2-xTB), SP2 and AG2 models, the EquiDTB1 and EquiDTB3 models, and the rMLP potential. The reference values for the energetic rankings were calculated using PBE0+MBD. All calculations consider a many-body dispersion treatment, except for xTB which considers D4 correction.

Dataset	metric	DFTB3	GFN2-xTB	SP2	AG2	EquiDTB3	EquiDTB1	rMLP
Equilibrium conformers	$\langle \mathrm{MAE}\ \Delta R_{\mathrm{K}} \rangle$	4.9	4.9	5.6	9.7	2.7	3.2	4.0
	$\langle \sigma \rangle$	4.3	4.1	4.6	7.1	2.6	2.9	3.5
	$\langle (\Delta R_{\mathrm{K}})_{max} \rangle$	16.7	16.1	18.1	26.8	10.4	12.2	14.2
Equilibrium and non-equilibrium conformers	$\langle \mathrm{MAE}\ \Delta R_{\mathrm{K}} \rangle$	0.53	0.31	0.31	0.48	0.12	0.18	0.21
	$\langle \sigma \rangle$	0.72	0.53	0.52	0.68	0.29	0.38	0.41
	$\langle (\Delta R_{\mathrm{K}})_{max} \rangle$	2.55	1.76	1.74	2.39	0.99	1.24	1.36