Supporting Information

T-ALPHA: a hierarchical transformer-based deep neural network for protein-ligand binding affinity prediction with uncertainty-aware self-learning for protein-specific alignment

Gregory W. Kyro, Anthony M. Smaldone, Yu Shee, Chuzhi Xu, Victor S. Batista

Department of Chemistry, Yale University

Corresponding author emails:
{gregory.kyro, victor.batista}@yale.edu

Contents

6 Protein-Specific Mpro and EGFR Benchmarks

- Figure S13. Predicted versus true binding affinities for T-ALPHA on the Mpro test set. Performance metrics include Root Mean Square Error (RMSE), Mean Absolute Error (MAE), coefficient of determination (r^2), Pearson correlation coefficient (r), and Spearman rank correlation coefficient (ρ).
- Figure S14. Predicted versus true binding affinities for T-ALPHA† on the Chai1-generated protein-ligand complex structures of the Mpro test set. Performance metrics include Root Mean Square Error (RMSE), Mean Absolute Error (MAE), coefficient of determination (r^2), Pearson correlation coefficient (r), and Spearman rank correlation coefficient (ρ).
- Figure S15. Predicted versus true binding affinities for T-ALPHA on the EGFR test set. Performance metrics include Root Mean Square Error (RMSE), Mean Absolute Error (MAE), coefficient of determination (r^2), Pearson correlation coefficient (r), and Spearman rank correlation coefficient (ρ).
- Figure S16. Predicted versus true binding affinities for T-ALPHA† on the Chai1-generated protein-ligand complex structures of the EGFR test set. Performance metrics include Root Mean Square Error (RMSE), Mean Absolute Error (MAE), coefficient of determination (r^2), Pearson correlation coefficient (r), and Spearman rank correlation coefficient (ρ).

7 Uncertainty-Aware Self-Learning Method

- Table S17. Improvements in Spearman rank correlation coefficient (ρ) for Mpro using the proposed self-learning method. Performance metrics are shown for comparing the baseline, new model, and fine-tuned model for SARS-CoV-2 main protease (Mpro). Spearman ρ improvements of 9.91% for the new model and 5.43% for the fine-tuned model highlight the effectiveness of the self-learning method. Results are reported separately for crystal structures (first value) and Chai1-generated structures (values in parentheses).
- Table S18. Improvements in Spearman rank correlation coefficient (ρ) for EGFR using the proposed self-learning method. Performance metrics are shown for comparing the baseline, new model, and fine-tuned model for epidermal growth factor receptor (EGFR). Spearman ρ improvements of 3.41% for the new model and 1.14% for the fine-tuned model highlight the effectiveness of the self-learning method. Results are reported separately for crystal structures (first value) and Chai1-generated structures (values in parentheses).

Table S1. List of all descriptors in the ligand physicochemical property-based feature vector.

MaxAbsEStateIndex	MaxEStateIndex	MinAbsEStateIndex	MinEStateIndex
qed	MolWt	HeavyAtomMolWt	ExactMolWt
NumValenceElectrons	NumRadicalElectrons	MaxPartialCharge	MinPartialCharge
MaxAbsPartialCharge	MinAbsPartialCharge	FpDensityMorgan1	FpDensityMorgan2
FpDensityMorgan3	BCUT2D_MWHI	BCUT2D_MWLOW	BCUT2D_CHGHI
BCUT2D_CHGLO	BCUT2D_LOGPHI	BCUT2D_LOGPLOW	BCUT2D_MRHI
BCUT2D_MRLOW	AvgIpc	BalabanJ	BertzCT
Chi0	Chi0n	Chi0v	Chi1
Chi1n	Chi1v	Chi2n	Chi2v
Chi3n	Chi3v	Chi4n	Chi4v
HallKierAlpha	Ipc	Kappa1	Kappa2
Kappa3	LabuteASA	PEOE_VSA1	PEOE_VSA10
PEOE_VSA11	PEOE_VSA12	PEOE_VSA13	PEOE_VSA14
PEOE_VSA2	PEOE_VSA3	PEOE_VSA4	PEOE_VSA5
PEOE_VSA6	PEOE_VSA7	PEOE_VSA8	PEOE_VSA9
SMR_VSA1	SMR_VSA10	SMR_VSA2	SMR_VSA3
SMR_VSA4	SMR_VSA5	SMR_VSA6	SMR_VSA7
SMR_VSA8	SMR_VSA9	SlogP_VSA1	SlogP_VSA10
SlogP_VSA11	SlogP_VSA12	SlogP_VSA2	SlogP_VSA3
SlogP_VSA4	SlogP_VSA5	SlogP_VSA6	SlogP_VSA7
SlogP_VSA8	SlogP_VSA9	TPSA	EState_VSA1
EState_VSA10	EState_VSA11	EState_VSA2	EState_VSA3
EState_VSA4	EState_VSA5	EState_VSA6	EState_VSA7
EState_VSA8	EState_VSA9	VSA_EState1	VSA_EState10
VSA_EState2	VSA_EState3	VSA_EState4	VSA_EState5
VSA_EState6	VSA_EState7	VSA_EState8	VSA_EState9
FractionCSP3	HeavyAtomCount	NHOHCount	NOCount
NumAliphaticCarbocycles	NumAliphaticHeterocycles	NumAliphaticRings	NumAromaticCarbocycles
NumAromaticHeterocycles	NumAromaticRings	NumHAcceptors	NumHDonors
NumHeteroatoms	NumRotatableBonds	NumSaturatedCarbocycles	NumSaturatedHeterocycles
NumSaturatedRings	RingCount	MolLogP	MolMR
fr_Al_COO	fr_Al_OH	fr_Al_OH_noTert	fr_ArN
fr_Ar_COO	fr_Ar_N	fr_Ar_NH	fr_Ar_OH
fr_COO	fr_COO2	fr_C_O	fr_C_O_noCOO
fr_C_S	fr_HOCCN	fr_Imine	fr_NH0
fr_NH1	fr_NH2	fr_N_O	fr_Ndealkylation1
fr_Ndealkylation2	fr_Nhpyrrole	fr_SH	fr_aldehyde
fr_alkyl_carbamate	fr_alkyl_halide	fr_allylic_oxid	fr_amide
fr_amidine	fr_aniline	fr_aryl_methyl	fr_azide
fr_azo	fr_barbitur	fr_benzene	fr_benzodiazepine
fr_bicyclic	fr_diazo	fr_dihydropyridine	fr_epoxide
fr_ester	fr_ether	fr_furan	fr_guanido
fr_halogen	fr_hdrzine	fr_hdrzone	fr_imidazole
fr_imide	fr_isocyan	fr_isothiocyan	fr_ketone
fr_ketone_Topliss	fr_lactam	fr_lactone	fr_methoxy
fr_morpholine	fr_nitrile	fr_nitro	fr_nitro_arom
fr_nitro_arom_nonortho	fr_nitroso	fr_oxazole	fr_oxime
fr_para_hydroxylation	fr_phenol	fr_phenol_noOrthoHbond	fr_phos_acid
fr_phos_ester	fr_piperdine	fr_piperzine	fr_priamide
fr_prisulfonamd	fr_pyridine	fr_quatN	fr_sulfide
fr_sulfonamd	fr_sulfone	fr_term_acetylene	fr_tetrazole
fr_thiazole	fr_thiocyan	fr_thiophene	fr_unbrch_alkane
fr_urea			

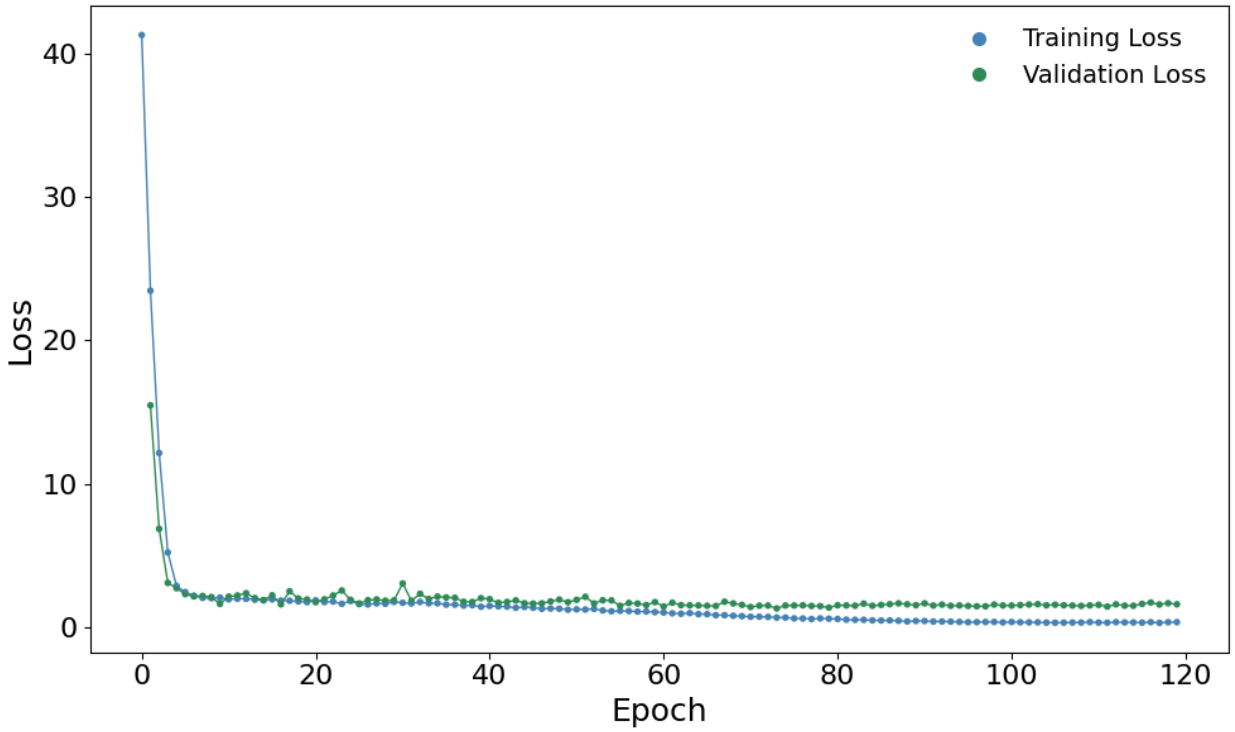

Figure S2. Training and validation loss curves for T-ALPHA corresponding to testing on the CASF 2016 test set. Loss is calculated using the custom loss function and plotted as a function of epoch number.

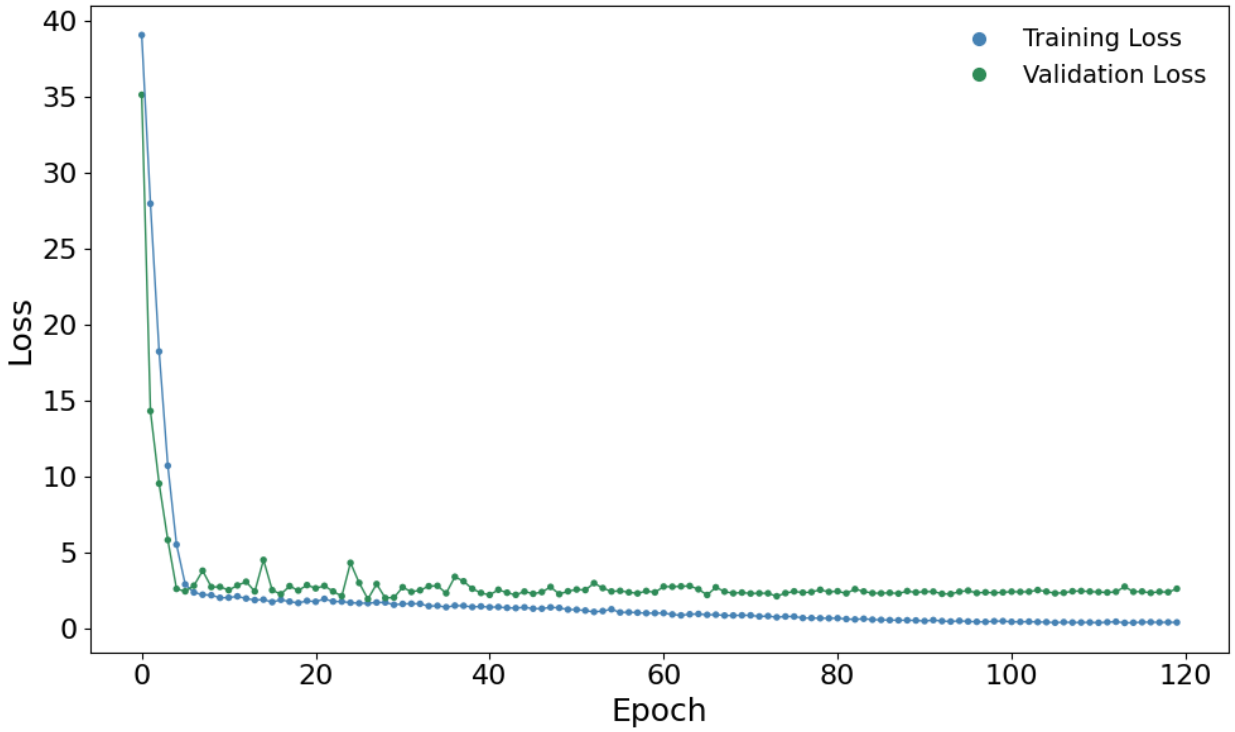

Figure S3. Training and validation loss curves for T-ALPHA corresponding to testing on the LP-PDBbind test set. Loss is calculated using the custom loss function and plotted as a function of epoch number.

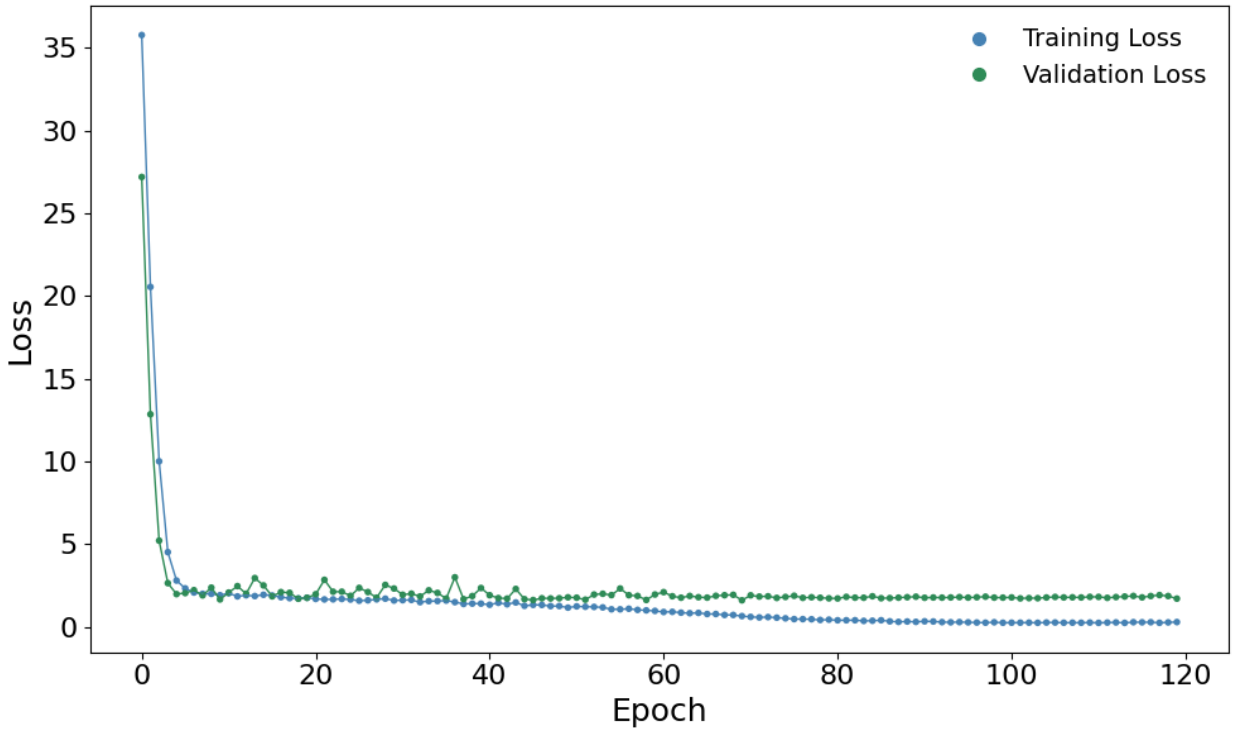

Figure S4. Training and validation loss curves for T-ALPHA corresponding to testing on the BDB2020+ test set, as well as protein-specific test sets for SARS-CoV-2 main protease (Mpro) and epidermal growth factor receptor (EGFR). Loss is calculated using the custom loss function and plotted as a function of epoch number.

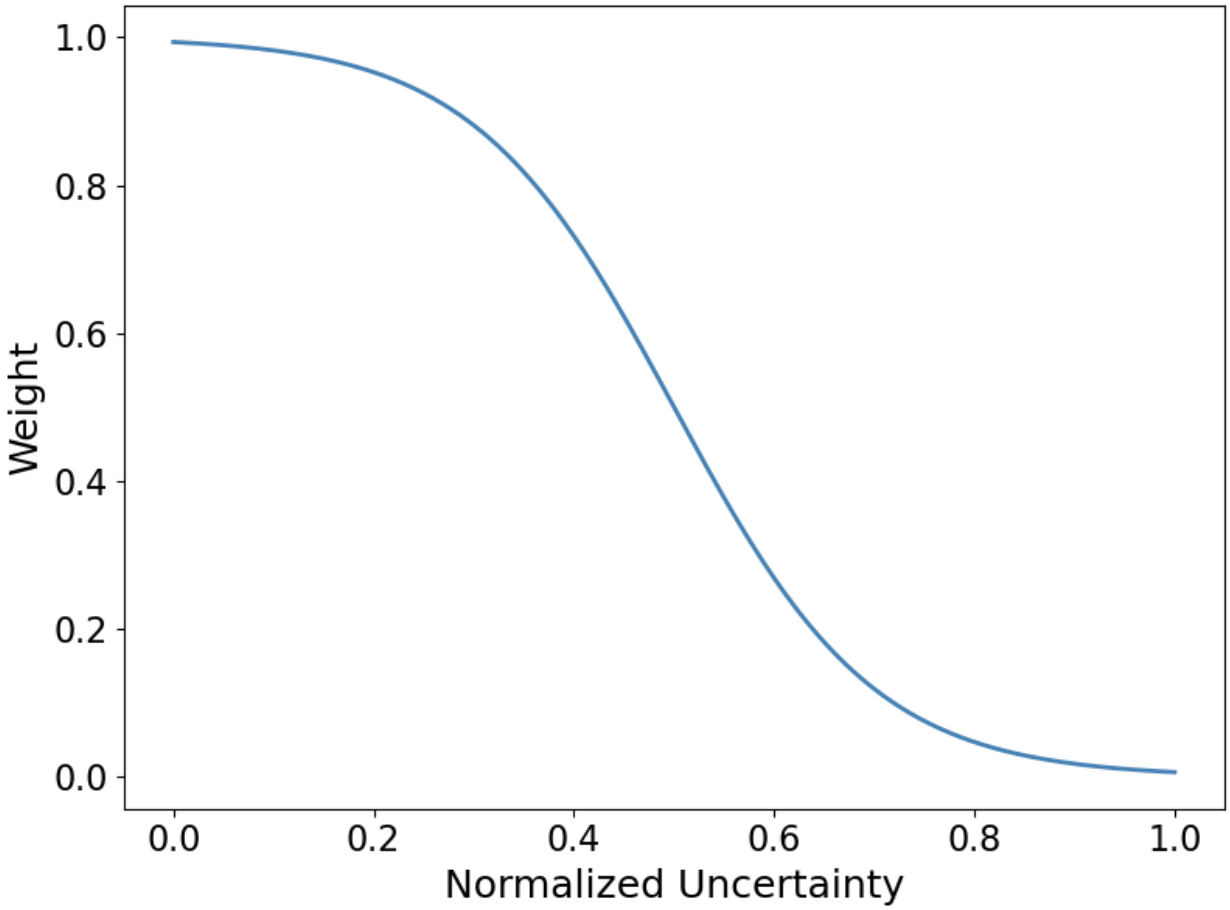

Figure S5. Weighting function for uncertainty-aware training. The figure shows the transformation applied to normalized uncertainty values to calculate weights for the loss function during parameter optimization. The function is defined as $w_i = 1 - \frac{1}{1+e^{-10(\sigma_i - 0.5)}}$, where σ_i is the normalized uncertainty. The transformation maps normalized uncertainty values to weights in the range (0, 1), smoothly decreasing the weight as uncertainty increases.

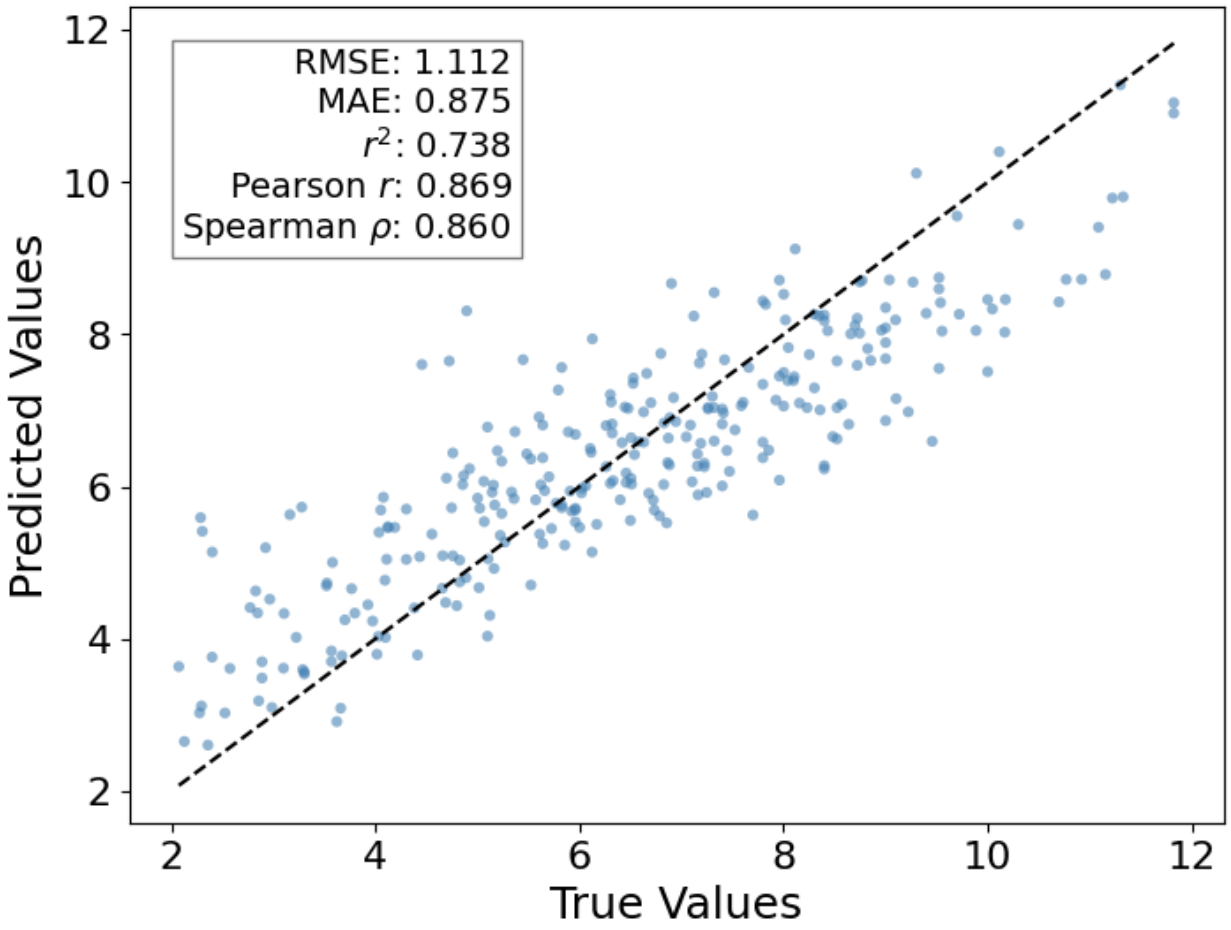

Figure S6. Predicted versus true binding affinities for T-ALPHA on the CASF 2016 test set using crystal structures of protein-ligand complexes. Performance metrics include Root Mean Square Error (RMSE), Mean Absolute Error (MAE), coefficient of determination (r^2), Pearson correlation coefficient (r), and Spearman rank correlation coefficient (ρ).

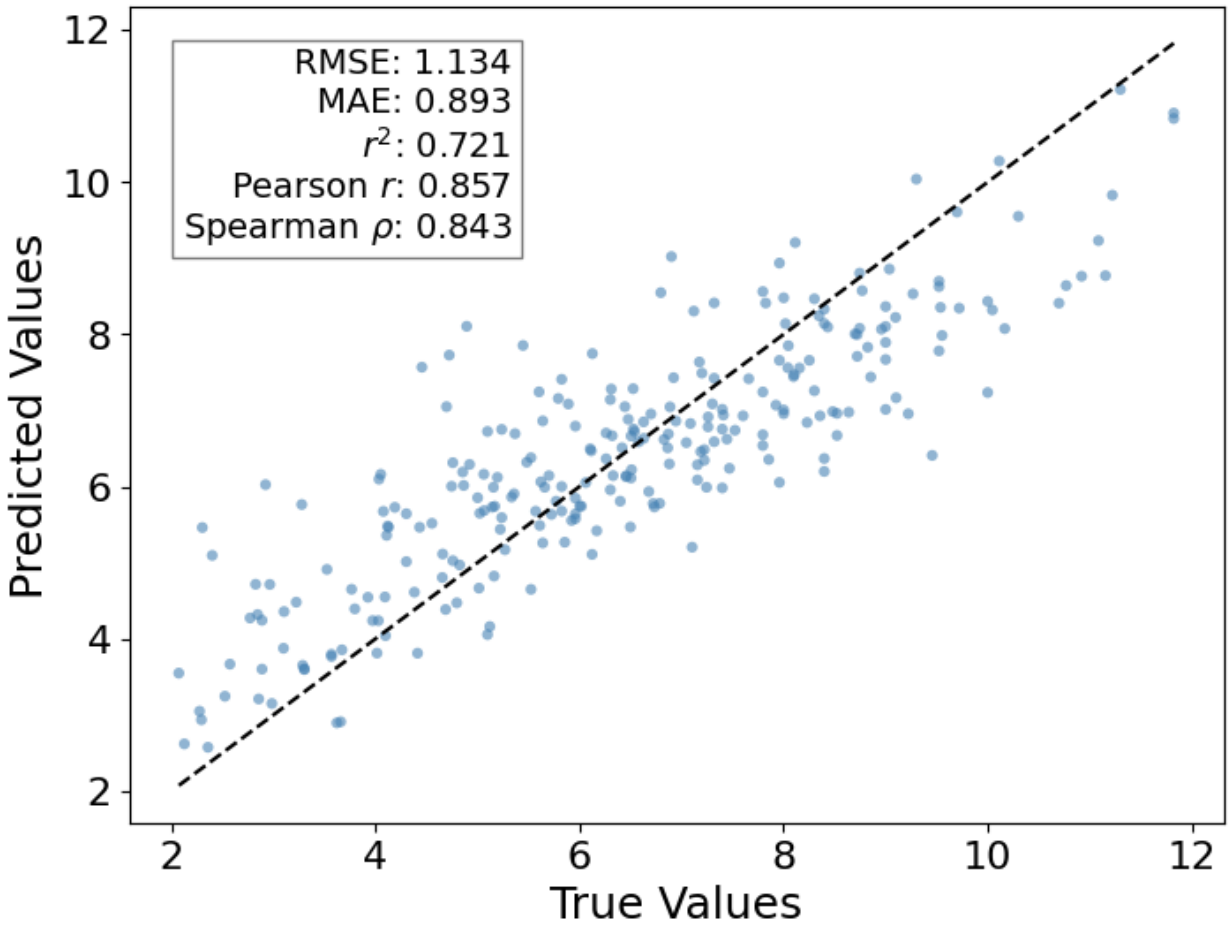

Figure S7. Predicted versus true binding affinities for T-ALPHA† on the CASF 2016 test set using Chai1-generated protein-ligand complex structures. Performance metrics include Root Mean Square Error (RMSE), Mean Absolute Error (MAE), coefficient of determination (r^2), Pearson correlation coefficient (r), and Spearman rank correlation coefficient (ρ).

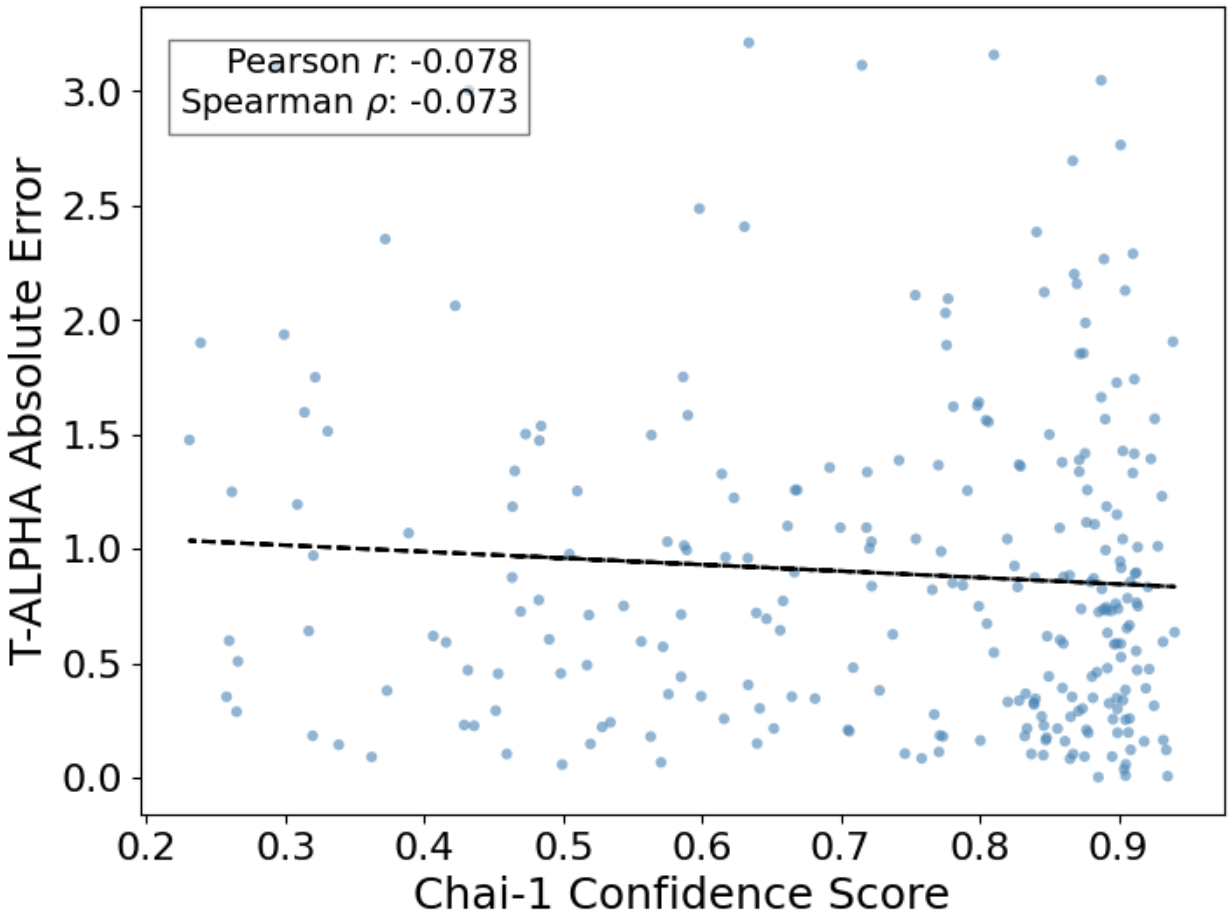

Figure S8. Absolute error of predicted binding affinities on the CASF 2016 test set using Chai1-generated protein-ligand complex structures as a function of the prediction confidence of Chai-1. Relevant metrics include Pearson correlation coefficient (r) and Spearman rank correlation coefficient (ρ).

Table S9. Performance of T-ALPHA and models reported in the literature on the LP-PDBbind test set.

Model	RMSE	MAE	r^2	Pearson r	Spearman ρ
T-ALPHA	**1.498**	**1.183**	**0.258**	**0.549**	**0.533**
AutoDock Vina	1.88	N/R	N/R	N/R	N/R
IGN	1.58	N/R	N/R	N/R	N/R
RF-Score	1.54	N/R	N/R	N/R	N/R
DeepDTA	1.68	N/R	N/R	N/R	N/R

[a] The table reports Root Mean Square Error (RMSE), Mean Absolute Error (MAE), coefficient of determination (r^2), Pearson correlation coefficient (r), and Spearman rank correlation coefficient (ρ) for each model.
[b] The best value for each metric is shown in bold.
[c] N/R indicates not reported in the literature.
[d] Error metrics (RMSE and MAE) are reported in units of pK_i / pK_d.

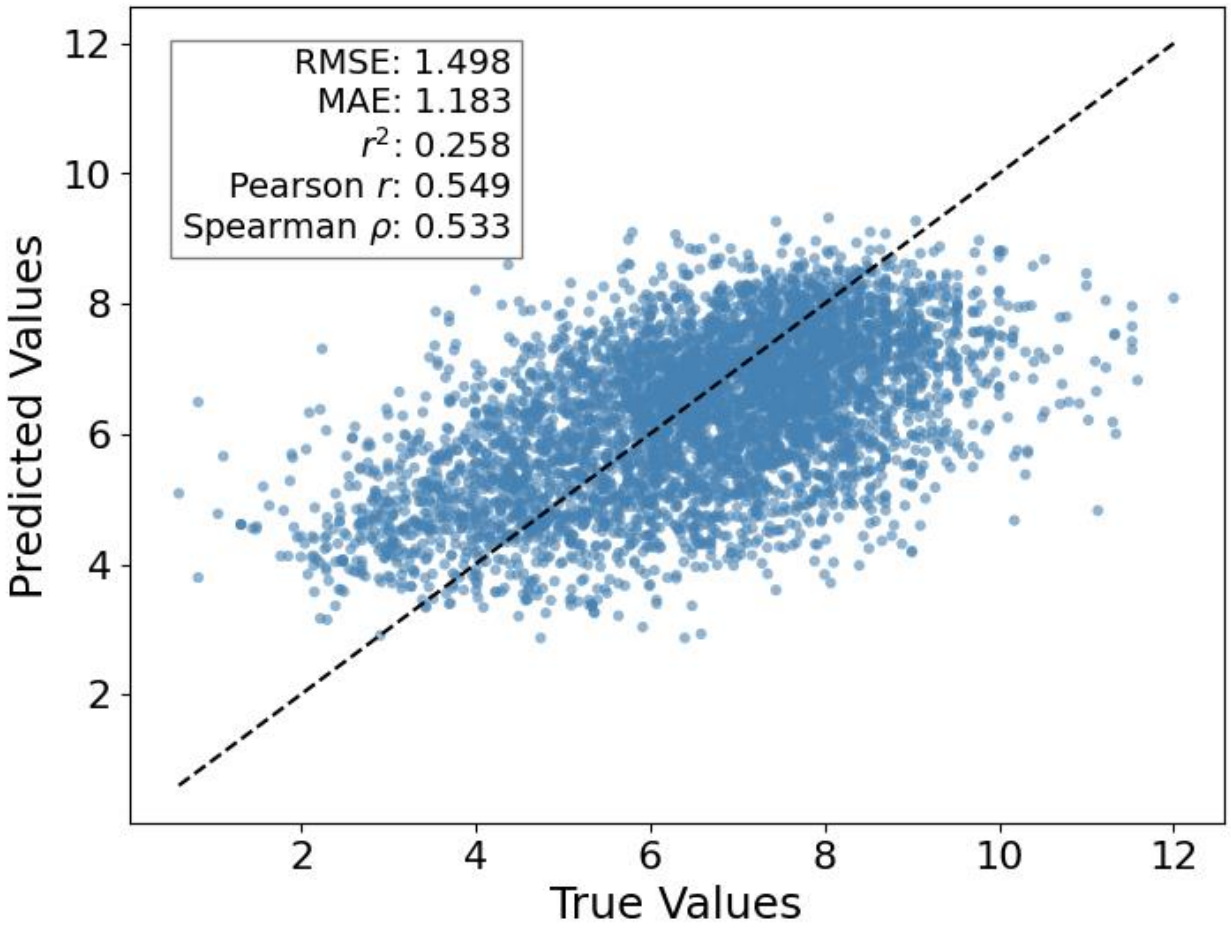

Figure S10. Predicted versus true binding affinities for T-ALPHA on the LP-PDBbind test set. Performance metrics include Root Mean Square Error (RMSE), Mean Absolute Error (MAE), coefficient of determination (r^2), Pearson correlation coefficient (r), and Spearman rank correlation coefficient (ρ).

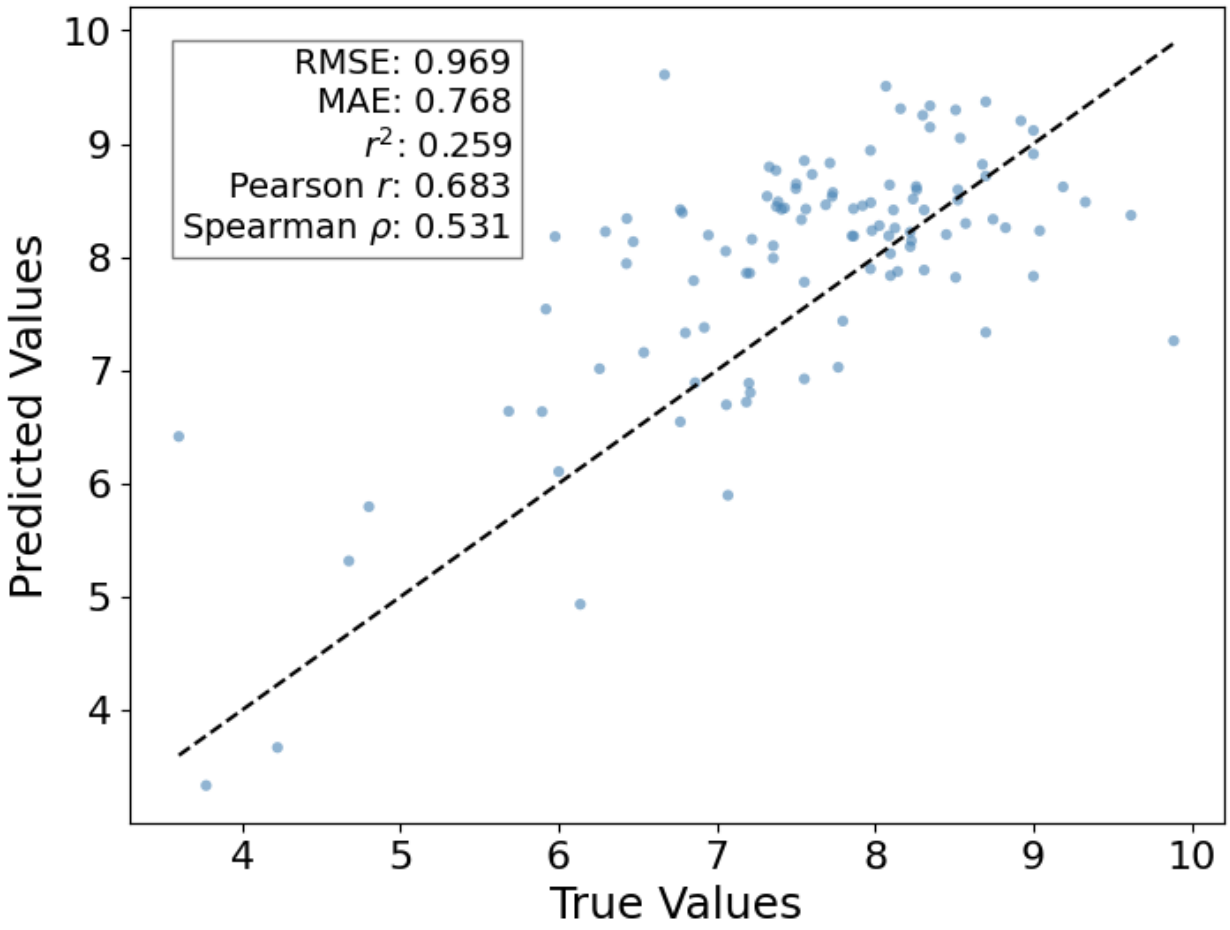

Figure S11. Predicted versus true binding affinities for T-ALPHA on the BDB2020+ test set. Performance metrics include Root Mean Square Error (RMSE), Mean Absolute Error (MAE), coefficient of determination (r^2), Pearson correlation coefficient (r), and Spearman rank correlation coefficient (ρ).

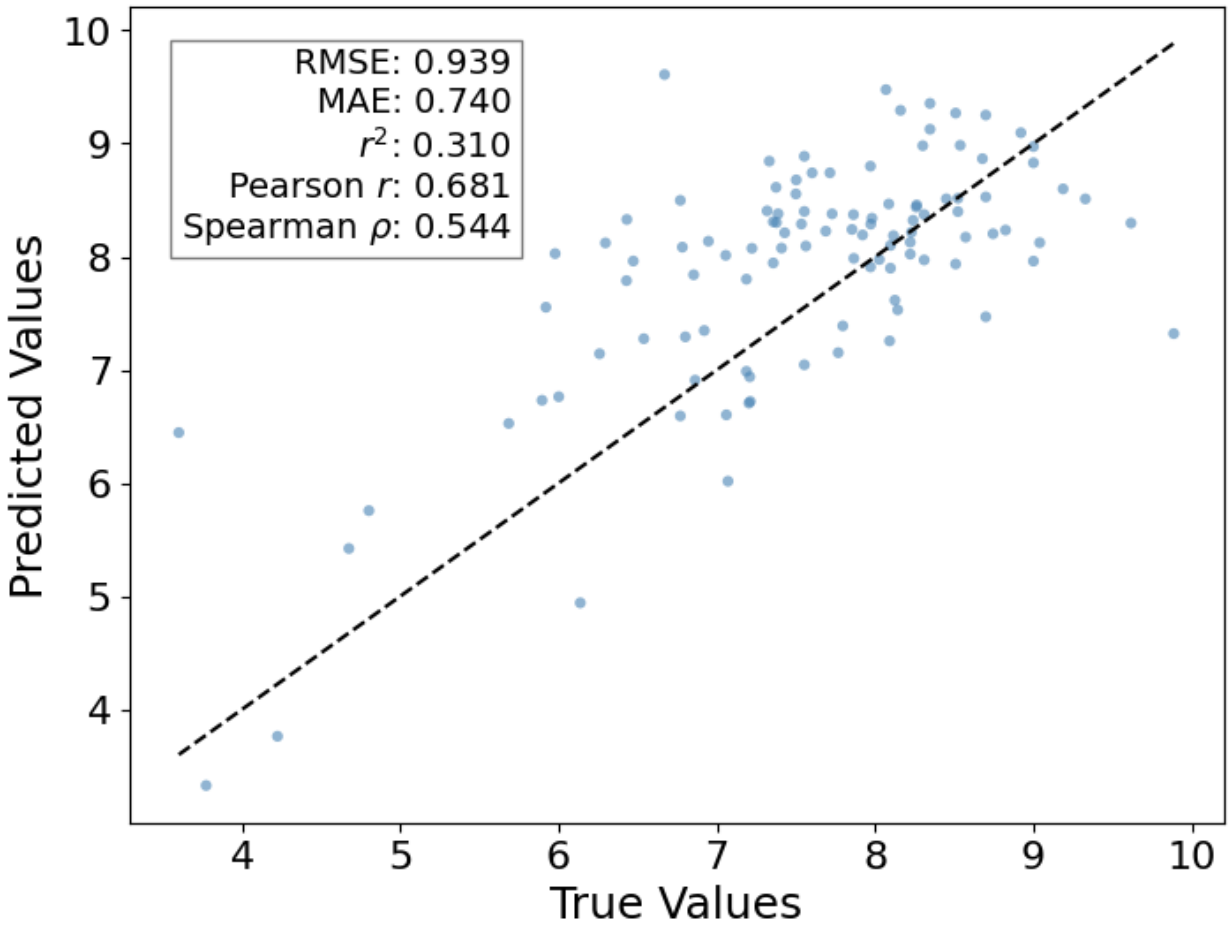

Figure S12. Predicted versus true binding affinities for T-ALPHA† on the Chai1-generated protein-ligand complex structures of the BDB2020+ test set. Performance metrics include Root Mean Square Error (RMSE), Mean Absolute Error (MAE), coefficient of determination (r^2), Pearson correlation coefficient (r), and Spearman rank correlation coefficient (ρ).

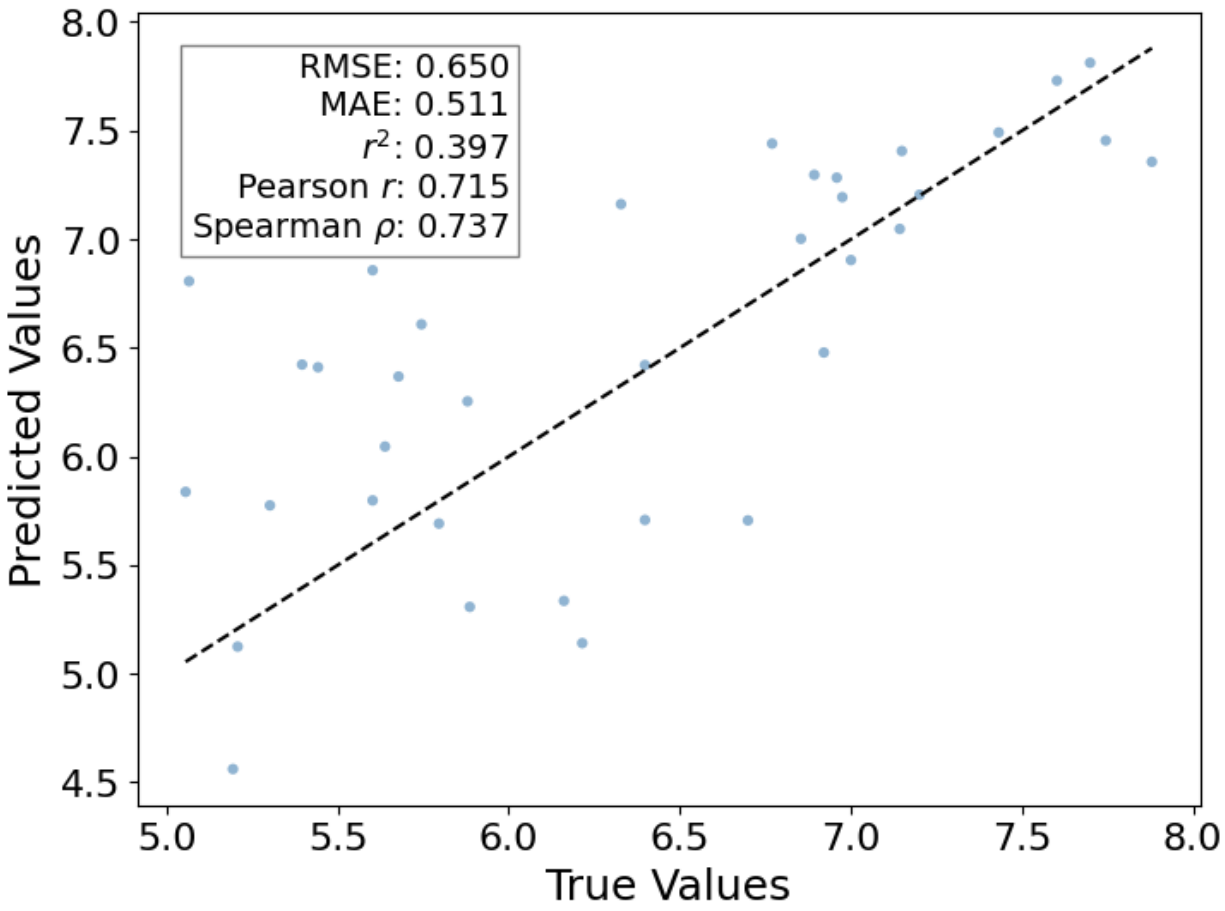

Figure S13. Predicted versus true binding affinities for T-ALPHA on the Mpro test set. Performance metrics include Root Mean Square Error (RMSE), Mean Absolute Error (MAE), coefficient of determination (r^2), Pearson correlation coefficient (r), and Spearman rank correlation coefficient (ρ).

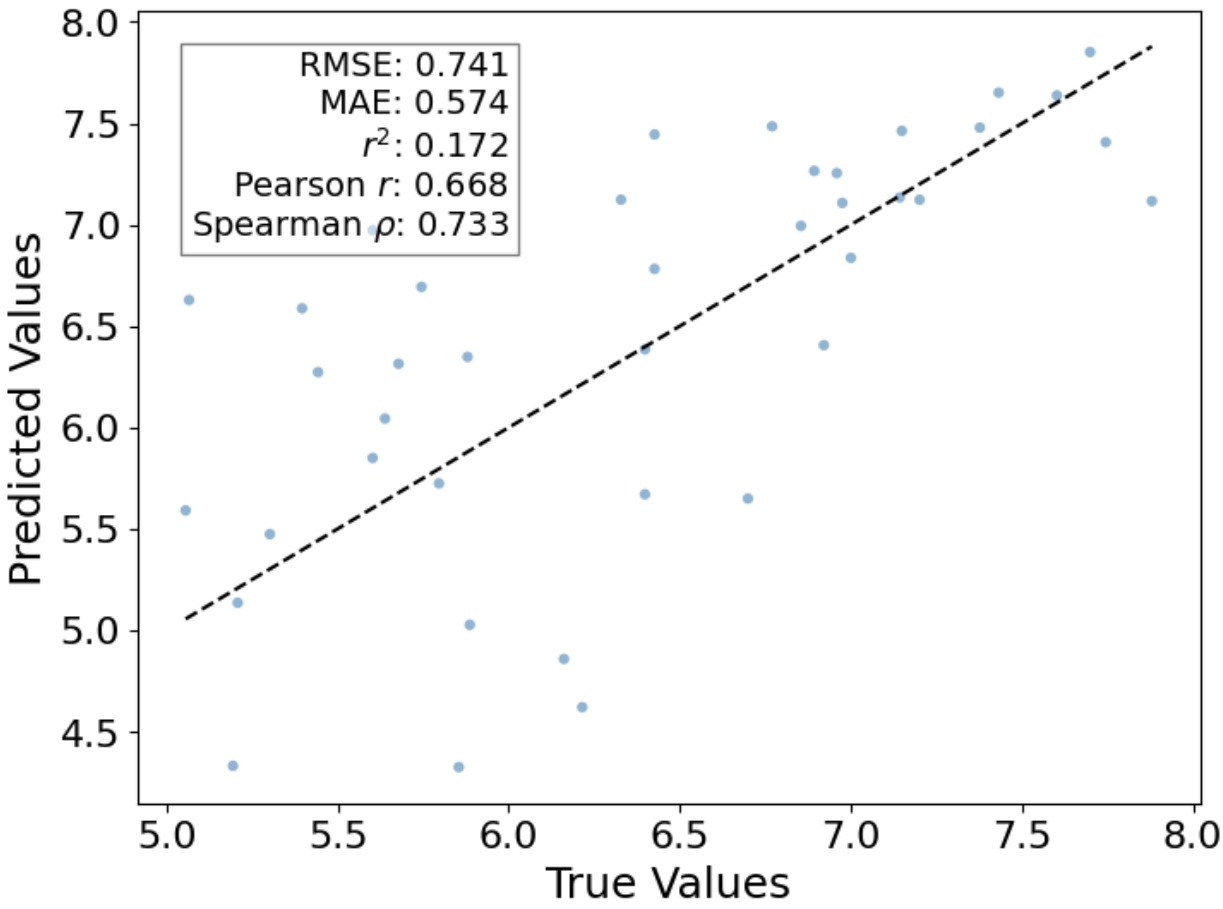

Figure S14. Predicted versus true binding affinities for T-ALPHA† on the Chai1-generated protein-ligand complex structures of the Mpro test set. Performance metrics include Root Mean Square Error (RMSE), Mean Absolute Error (MAE), coefficient of determination (r^2), Pearson correlation coefficient (r), and Spearman rank correlation coefficient (ρ).

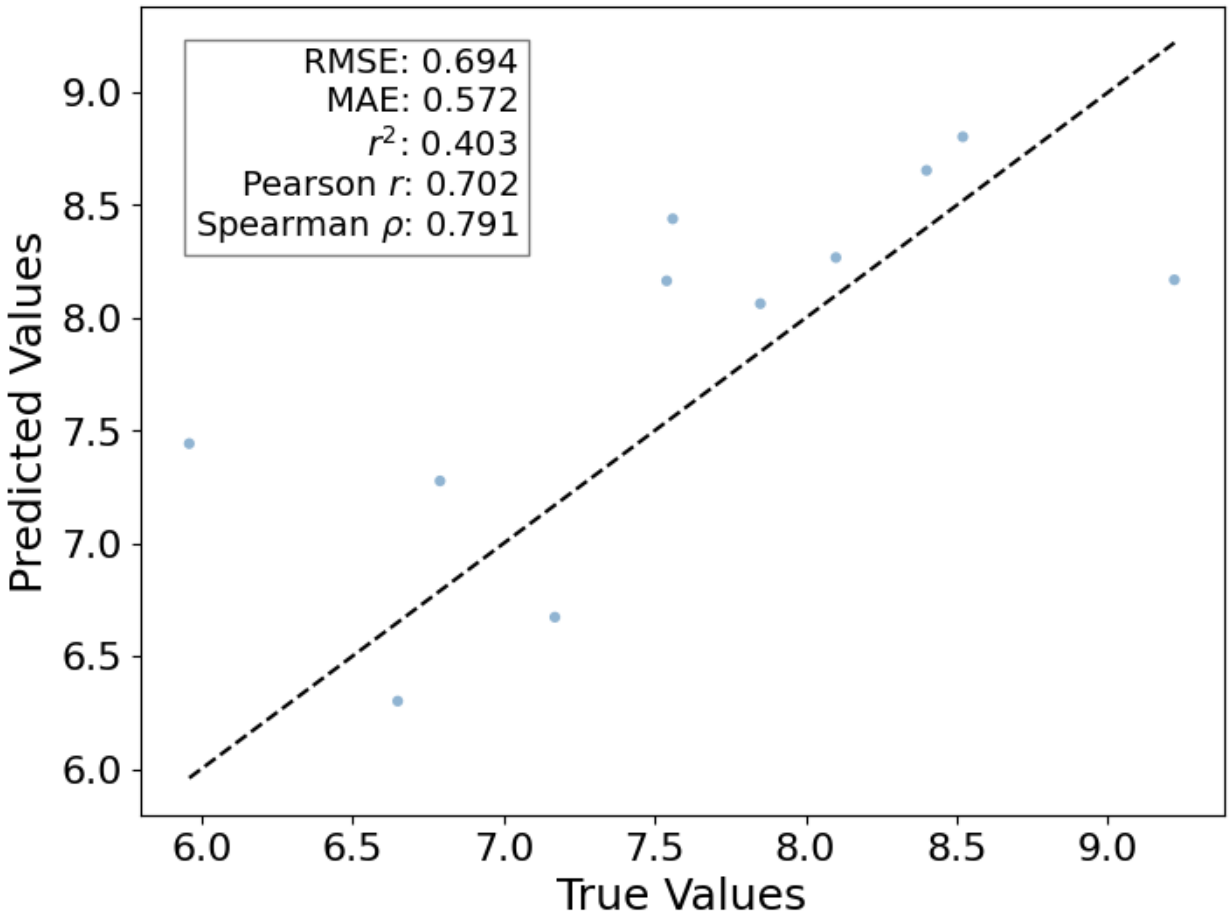

Figure S15. Predicted versus true binding affinities for T-ALPHA on the EGFR test set. Performance metrics include Root Mean Square Error (RMSE), Mean Absolute Error (MAE), coefficient of determination (r^2), Pearson correlation coefficient (r), and Spearman rank correlation coefficient (ρ).

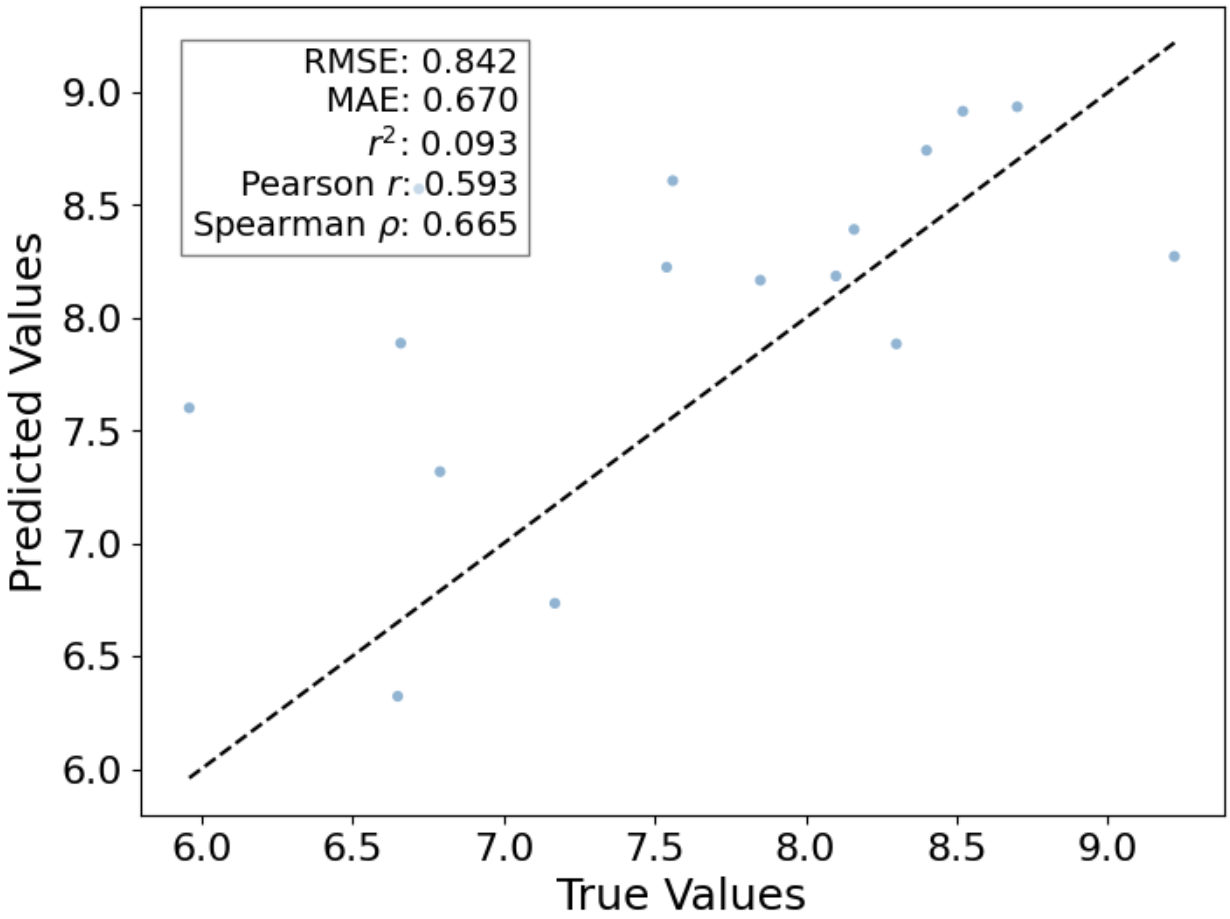

Figure S16. Predicted versus true binding affinities for T-ALPHA† on the Chai1-generated protein-ligand complex structures of the EGFR test set. Performance metrics include Root Mean Square Error (RMSE), Mean Absolute Error (MAE), coefficient of determination (r^2), Pearson correlation coefficient (r), and Spearman rank correlation coefficient (ρ).

Table S17. Improvements in Spearman rank correlation coefficient (ρ) for Mpro using the proposed self-learning method. Performance metrics are shown for comparing the baseline, new model, and fine-tuned model for SARS-CoV-2 main protease (Mpro). Spearman ρ improvements of 9.91% for the new model and 5.43% for the fine-tuned model highlight the effectiveness of the self-learning method. Results are reported separately for crystal structures (first value) and Chai1-generated structures (values in parentheses).

Model	RMSE	MAE	Pearson *r*	Spearman ρ	% ↑ Δ Spearman ρ
Baseline	0.650 (0.741)	0.511 (0.574)	0.715 (0.668)	0.737 (0.733)	0.00% (0.00%)
New Model	1.397 (1.468)	1.259 (1.324)	0.790 (0.776)	0.810 (0.803)	9.91% (9.55%)
Control (New Model)	1.374 (1.387)	1.210 (1.200)	0.741 (0.700)	0.737 (0.702)	0.00% (-4.22%)
Fine Tune	0.615 (0.651)	0.529 (0.549)	0.752 (0.718)	0.777 (0.762)	5.43% (3.96%)
Control (Fine Tune)	0.915 (0.923)	0.779 (0.776)	0.764 (0.750)	0.735 (0.732)	-0.27% (-0.14%)

[a] The table reports Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Pearson correlation coefficient (*r*), Spearman rank correlation coefficient (ρ), and the percentage increase in Spearman ρ compared to the baseline (% ↑ Δ Spearman ρ).
[b] Error metrics (RMSE and MAE) are reported in units of pK_i / pK_d.

Table S18. Improvements in Spearman rank correlation coefficient (ρ) for EGFR using the proposed self-learning method. Performance metrics are shown for comparing the baseline, new model, and fine-tuned model for epidermal growth factor receptor (EGFR). Spearman ρ improvements of 3.41% for the new model and 1.14% for the fine-tuned model highlight the effectiveness of the self-learning method. Results are reported separately for crystal structures (first value) and Chai1-generated structures (values in parentheses).

Model	RMSE	MAE	Pearson *r*	Spearman ρ	% ↑ Δ Spearman ρ
Baseline	0.694 (0.842)	0.572 (0.670)	0.702 (0.593)	0.791 (0.665)	0.00% (0.00%)
New Model	3.132 (3.407)	3.078 (3.334)	0.772 (0.685)	0.818 (0.679)	3.41% (2.11%)
Control (New Model)	3.520 (3.507)	3.455 (3.432)	0.663 (0.580)	0.791 (0.668)	0.00% (0.45%)
Fine Tune	1.033 (1.010)	0.911 (0.858)	0.755 (0.665)	0.800 (0.709)	1.14% (6.62%)
Control (Fine Tune)	1.172 (1.103)	1.001 (0.900)	0.569 (0.545)	0.773 (0.579)	-2.28% (-12.93%)

[a] The table reports Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Pearson correlation coefficient (*r*), Spearman rank correlation coefficient (ρ), and the percentage increase in Spearman ρ compared to the baseline (% ↑ Δ Spearman ρ).
[b] Error metrics (RMSE and MAE) are reported in units of pK_i / pK_d.