Supplementary Information

A z-averaged ensemble of three orthogonal models improves protein–ligand virtual screening on leakage-controlled benchmarks

Yuguang Mu, Yang Liu, and Weifeng Li

Corresponding authors: Yuguang Mu, Weifeng Li. Email: ygmu@ntu.edu.sg; lwf@mail.sdu.edu.cn

---

This PDF file includes:

- Supplementary Note 1 (external validation on the MUV benchmark)
- Supplementary Fig. 1
- Supplementary Tables 1 to 5
- SI References

Supplementary Table 1. BIND-full as an industry reference baseline (not zero-shot)

BIND-full was fine-tuned with per-target actives present in each benchmark's test set; we report it for completeness but it is **not directly comparable** to the zero-shot models in Figure 2.

Benchmark	n	BIND-full AUROC	3-way ens. AUROC (zero-shot)	Δ (BIND-full − ens3)
DUD-E_uniprot full	102	0.945	0.858	+0.087
DUD-AD_uniprot full	102	0.927	0.811	+0.117
DUD-E-strict	7	0.946	0.779	+0.167
DUD-AD-strict	7	0.930	0.755	+0.175
bdb_strict_test	33	—	pending	—
LIT-PCBA	14	0.609	0.608	+0.001

BIND-full's apparent advantage on DUD-E/DUD-AD subsets is driven by in-distribution memorization of test-target actives, not by superior modeling. Under the leakage-clean bdb_strict_test split, BIND-full was not run because the training data contains the strict-test targets by construction.

Supplementary Table 2. Per-target AUROC across six benchmarks

DUD-E (102)

Target	BIND-z	BIND-full	SaProt-v2	DrugCLIP	ens2	ens3
aa2ar	0.9520	0.9948	0.7663	0.8855	0.9568	0.9684
abl1	0.9288	0.9720	0.7752	0.4491	0.9304	0.8834
ace	0.8840	0.9914	0.8631	0.4103	0.9201	0.8518
aces	0.9742	0.9707	0.3628	0.9242	0.9074	0.9625
ada	0.9954	0.9976	0.6437	0.8385	0.9823	0.9913
ada17	0.8484	0.9813	0.9714	0.4536	0.9459	0.9333
adrb1	0.9739	0.9678	0.7525	0.7817	0.9594	0.9391
adrb2	0.9628	0.9514	0.7548	0.6910	0.9517	0.9633
akt1	0.9357	0.9546	0.7853	0.7318	0.9135	0.8776
akt2	0.9462	0.9692	0.7437	0.6321	0.9390	0.9190
aldr	0.8791	0.9559	0.3517	0.4529	0.6890	0.6273
ampc	0.7733	0.8785	0.8081	0.6266	0.8465	0.8477
andr	0.5674	0.9295	0.7430	0.6484	0.7200	0.7397
aofb	0.8656	0.8795	0.4987	0.5099	0.7411	0.6925
bace1	0.9343	0.9722	0.7306	0.7645	0.9105	0.9111
braf	0.9506	0.9812	0.8520	0.6198	0.9372	0.8901
cah2	0.9940	0.9958	0.9553	0.3225	0.9942	0.9787

casp3	0.9205	0.9205	0.7134	0.8209	0.8878	0.9318
cdk2	0.8994	0.9121	0.7670	0.6259	0.9102	0.8769
comt	0.7912	0.9933	0.8343	0.8887	0.8764	0.9689
cp2c9	0.5072	0.8042	0.5528	0.5126	0.5423	0.5457
cp3a4	0.5730	0.8180	0.5594	0.3852	0.5633	0.5320
csf1r	0.9223	0.9764	0.8069	0.6298	0.9182	0.9162
cxcr4	0.4039	0.9093	0.7002	0.5801	0.5956	0.5860
def	0.9967	0.9997	0.8613	0.8158	0.9974	0.9981
dhi1	0.8933	0.9726	0.5274	0.4808	0.8386	0.8061
dpp4	0.9310	0.9855	0.8698	0.3102	0.9424	0.9065
drd3	0.9731	0.9896	0.9248	0.8329	0.9746	0.9719
dyr	0.9837	0.9941	0.9446	0.9773	0.9865	0.9896
egfr	0.9285	0.9852	0.8380	0.6361	0.9380	0.9389
esr1	0.9214	0.9903	0.8753	0.9409	0.9315	0.9531
esr2	0.8861	0.9894	0.7887	0.9603	0.9042	0.9512
fa10	0.8884	0.9938	0.8169	0.6213	0.9041	0.9110
fa7	0.9039	0.9744	0.8829	0.3417	0.9386	0.9027
fabp4	0.9320	0.9720	0.5345	0.8136	0.8587	0.9041
fak1	0.9878	0.9570	0.9653	0.5478	0.9972	0.9906
fgfr1	0.9704	0.9723	0.9129	0.4334	0.9748	0.9616

fkb1a	0.8799	0.9797	0.7230	0.9780	0.8803	0.9746
fnta	0.9941	0.9966	0.5377	0.9079	0.9796	0.9903
fpps	0.9994	1.0000	0.9995	0.8498	0.9999	0.9996
gcr	0.5378	0.9675	0.8615	0.7174	0.7903	0.7980
glcm	0.7985	0.9192	0.8180	0.5219	0.8016	0.8158
gria2	0.8981	0.8346	0.3175	0.6101	0.6805	0.6847
grik1	0.9484	0.9159	0.2387	0.9240	0.7315	0.8747
hdac2	0.9984	0.9925	0.8614	0.9554	0.9985	0.9960
hdac8	0.9979	0.9935	0.9620	0.9634	0.9976	0.9953
hivint	0.4769	0.6722	0.4212	0.8652	0.4436	0.7049
hivpr	0.3693	0.6985	0.7196	0.6050	0.6008	0.6447
hivrt	0.4277	0.4816	0.6040	0.2520	0.5380	0.3582
hmdh	0.6993	0.9996	0.3889	0.5652	0.5848	0.6008
hs90a	0.9710	0.9977	0.7578	0.9044	0.9711	0.9849
hxk4	0.4643	0.9966	0.1670	0.6563	0.2443	0.3503
igf1r	0.9226	0.9874	0.9250	0.6283	0.9600	0.9557
inha	0.8578	0.9237	0.7625	0.6068	0.8525	0.7789
ital	0.5738	0.9002	0.5798	0.5726	0.6343	0.6408
jak2	0.9117	0.9841	0.9073	0.5974	0.9671	0.9497
kif11	0.5223	0.9333	0.4905	0.5974	0.5126	0.5516

kit	0.9209	0.9847	0.8963	0.6704	0.9545	0.9392
kith	0.8264	0.8176	0.3554	0.5167	0.6981	0.6734
kpcb	0.9534	0.9557	0.9194	0.8170	0.9720	0.9739
lck	0.9615	0.9862	0.8829	0.5070	0.9662	0.9404
lkha4	0.3891	0.9917	0.6156	0.7707	0.4970	0.6744
mapk2	0.8620	0.9662	0.8919	0.6061	0.9064	0.8736
mcr	0.6484	0.9415	0.8104	0.6595	0.7936	0.8096
met	0.8250	0.9844	0.7677	0.1715	0.8610	0.7597
mk01	0.8024	0.9292	0.8840	0.7203	0.9252	0.9336
mk10	0.7919	0.9304	0.7174	0.6939	0.8022	0.8429
mk14	0.8929	0.9710	0.6277	0.7202	0.8651	0.9023
mmp13	0.9676	0.9956	0.9430	0.2341	0.9819	0.9662
mp2k1	0.9183	0.9619	0.7874	0.6556	0.9280	0.9257
nos1	0.7850	0.9097	0.5550	0.8338	0.7353	0.8362
nram	0.9634	0.9953	0.7090	0.9877	0.9495	0.9911
pa2ga	0.8046	0.7974	0.4076	0.7221	0.6651	0.7606
parp1	0.7043	0.9639	0.8876	0.9337	0.8800	0.9618
pde5a	0.9007	0.9591	0.7727	0.7584	0.8954	0.9134
pgh1	0.7993	0.9003	0.5633	0.6119	0.7599	0.7600
pgh2	0.9032	0.9334	0.5534	0.6612	0.8656	0.8632

plk1	0.9078	0.9101	0.8465	0.4280	0.9245	0.8909
pnph	0.9033	0.9577	0.7856	0.9840	0.9296	0.9884
ppara	0.5192	0.9873	0.9362	0.9705	0.8321	0.9550
ppard	0.7298	0.9890	0.9582	0.9587	0.9106	0.9656
pparg	0.4531	0.9493	0.8309	0.9328	0.7185	0.9047
prgr	0.6258	0.9474	0.6627	0.7642	0.6692	0.7789
ptn1	0.9171	0.9194	0.5707	0.8857	0.8005	0.9016
pur2	0.8627	0.9849	0.6895	0.9280	0.8711	0.9639
pygm	0.5614	0.9767	0.4279	0.6579	0.5179	0.5864
pyrd	0.7546	0.9698	0.4868	0.4974	0.6596	0.6131
reni	0.8899	0.9456	0.7792	0.5514	0.8834	0.9043
rock1	0.9440	0.9847	0.8090	0.7410	0.9386	0.9770
rxra	0.8591	0.9739	0.9706	0.7747	0.9429	0.9395
sahh	0.5749	0.9978	0.7829	0.4559	0.7546	0.6722
src	0.9571	0.9866	0.6893	0.6641	0.9337	0.9308
tgfr1	0.9786	0.9990	0.6346	0.7814	0.9363	0.9654
thb	0.8719	0.9579	0.7569	0.6948	0.8665	0.8691
thrb	0.7001	0.9795	0.8292	0.5313	0.8027	0.7840
try1	0.9660	0.9679	0.9177	0.4075	0.9654	0.9459
tryb1	0.9337	0.9619	0.7899	0.7141	0.9389	0.9328

tysy	0.7650	0.8301	0.5884	0.8639	0.7159	0.8485
urok	0.9492	0.9849	0.7885	0.5531	0.9541	0.9345
vgfr2	0.9660	0.9841	0.8741	0.2937	0.9785	0.8921
wee1	0.9595	0.7881	0.7001	0.6002	0.9362	0.9352
xiap	0.9956	0.9842	0.7188	0.3643	0.9752	0.9538

26 *DUD-AD (102)*

Target	BIND-z	BIND-full	SaProt-v2	DrugCLIP	ens2	ens3
aa2ar	0.9434	0.9923	0.7354	0.8620	0.9385	0.9523
abl1	0.8963	0.9519	0.6210	0.5183	0.8469	0.8099
ace	0.8054	0.9952	0.9329	0.6255	0.9563	0.9382
aces	0.9763	0.9656	0.5591	0.9072	0.9397	0.9668
ada	0.9899	0.9902	0.7144	0.8508	0.9729	0.9851
ada17	0.8623	0.9825	0.9590	0.4892	0.9443	0.9322
adrb1	0.9844	0.9868	0.8290	0.8586	0.9765	0.9683
adrb2	0.9825	0.9887	0.8088	0.7859	0.9742	0.9819
akt1	0.8730	0.9412	0.7628	0.7699	0.8594	0.8547
akt2	0.8263	0.9458	0.6929	0.6901	0.8326	0.8459
aldr	0.8544	0.9354	0.5545	0.2981	0.7831	0.6249
ampc	0.6446	0.7477	0.7438	0.6794	0.7384	0.7754

andr	0.6750	0.8622	0.5376	0.6622	0.6449	0.6901
aofb	0.8171	0.8589	0.6094	0.4801	0.7731	0.6919
bace1	0.9010	0.9692	0.6402	0.8144	0.8534	0.8916
braf	0.8474	0.9532	0.6090	0.5938	0.7678	0.7262
cah2	0.9894	0.9863	0.8742	0.3912	0.9836	0.9548
casp3	0.9312	0.9380	0.6799	0.8046	0.8933	0.9319
cdk2	0.8080	0.8605	0.5685	0.6759	0.7603	0.7803
comt	0.7889	0.9482	0.6870	0.8673	0.8215	0.9208
cp2c9	0.4806	0.6587	0.5459	0.4565	0.5238	0.4979
cp3a4	0.5779	0.7567	0.6253	0.4193	0.6056	0.5749
csf1r	0.8339	0.9502	0.7186	0.6320	0.8262	0.8236
cxcr4	0.4515	0.8049	0.7188	0.5649	0.6429	0.6369
def	0.9905	0.9968	0.8095	0.8325	0.9892	0.9912
dhi1	0.8215	0.9597	0.5384	0.5708	0.7674	0.7697
dpp4	0.9444	0.9943	0.9196	0.3252	0.9628	0.9337
drd3	0.9746	0.9929	0.9441	0.8273	0.9804	0.9781
dyr	0.9854	0.9962	0.9488	0.9578	0.9879	0.9884
egfr	0.8920	0.9784	0.6002	0.5899	0.8355	0.8238
esr1	0.7770	0.9760	0.6012	0.9177	0.7323	0.8473
esr2	0.7654	0.9764	0.4420	0.9433	0.6594	0.8413

fa10	0.8592	0.9883	0.6826	0.5159	0.8207	0.8172
fa7	0.8112	0.9302	0.6544	0.2013	0.7883	0.6150
fabp4	0.9575	0.9972	0.6669	0.8881	0.9184	0.9705
fak1	0.9868	0.9721	0.8357	0.6985	0.9811	0.9767
fgfr1	0.9241	0.9218	0.7475	0.5686	0.9033	0.8864
fkb1a	0.7466	0.9542	0.5356	0.9659	0.6901	0.9044
fnta	0.9844	0.9891	0.5378	0.8791	0.9506	0.9687
fpps	0.7790	0.9323	0.4345	0.5109	0.6390	0.6106
gcr	0.5731	0.9095	0.6338	0.7299	0.6425	0.7086
glcm	0.7172	0.8420	0.7780	0.5197	0.7565	0.7340
gria2	0.8482	0.8627	0.4236	0.5067	0.6872	0.6458
grik1	0.9565	0.9298	0.2890	0.9342	0.8153	0.9145
hdac2	0.9946	0.9848	0.8240	0.9464	0.9899	0.9892
hdac8	0.9904	0.9854	0.9359	0.9580	0.9886	0.9890
hivint	0.5787	0.7137	0.3581	0.7649	0.4529	0.6280
hivpr	0.4723	0.7528	0.5310	0.6385	0.5236	0.6077
hivrt	0.5116	0.5392	0.5088	0.3453	0.5260	0.4202
hmdh	0.5903	0.9914	0.5219	0.3349	0.5887	0.4837
hs90a	0.9454	0.9921	0.6180	0.8609	0.9226	0.9531
hxk4	0.4924	0.9783	0.3710	0.6055	0.4022	0.4684

igf1r	0.7896	0.9664	0.8115	0.7448	0.8588	0.9109
inha	0.7608	0.8660	0.6570	0.6498	0.7385	0.7144
ital	0.6228	0.8954	0.5708	0.3485	0.6564	0.5444
jak2	0.8001	0.9161	0.7066	0.6189	0.8367	0.8456
kif11	0.5970	0.9335	0.4273	0.5669	0.5380	0.5599
kit	0.8500	0.9565	0.8013	0.5960	0.8885	0.8565
kith	0.7738	0.7916	0.5800	0.5837	0.7624	0.7386
kpcb	0.9140	0.9229	0.9167	0.7593	0.9585	0.9544
lck	0.9003	0.9712	0.6345	0.5914	0.8485	0.8334
lkha4	0.3664	0.9967	0.6753	0.5536	0.5119	0.5807
mapk2	0.7671	0.9263	0.7964	0.6745	0.8110	0.8118
mcr	0.6200	0.8470	0.5308	0.7059	0.5930	0.6765
met	0.7053	0.9759	0.7162	0.2758	0.7665	0.6787
mk01	0.6345	0.8896	0.7885	0.7816	0.7946	0.8756
mk10	0.7659	0.9206	0.5824	0.6925	0.7271	0.7953
mk14	0.7864	0.9198	0.5962	0.7273	0.7695	0.8441
mmp13	0.9608	0.9939	0.9311	0.2505	0.9773	0.9637
mp2k1	0.9046	0.9528	0.6082	0.5665	0.8769	0.8591
nos1	0.6846	0.8896	0.5371	0.6301	0.6445	0.6713
nram	0.9681	0.9576	0.5896	0.9832	0.9363	0.9860

pa2ga	0.8259	0.7994	0.5786	0.7315	0.7770	0.8257
parp1	0.5360	0.9600	0.8207	0.9253	0.7446	0.9135
pde5a	0.8273	0.9406	0.7208	0.7484	0.8296	0.8654
pgh1	0.7663	0.8948	0.5195	0.6336	0.7004	0.7204
pgh2	0.8786	0.9260	0.5614	0.6250	0.8415	0.8285
plk1	0.8570	0.9305	0.7796	0.4453	0.8714	0.8302
pnph	0.7517	0.9407	0.7662	0.9768	0.8267	0.9671
ppara	0.5240	0.9939	0.7037	0.9597	0.6420	0.8902
ppard	0.5722	0.9845	0.7851	0.9024	0.7154	0.8573
pparg	0.4804	0.9387	0.6843	0.9075	0.6070	0.8505
prgr	0.6394	0.9126	0.4836	0.7995	0.5733	0.7369
ptn1	0.9299	0.8840	0.7226	0.8453	0.8816	0.9073
pur2	0.9039	0.9909	0.5406	0.9825	0.8506	0.9785
pygm	0.5659	0.9869	0.6368	0.6221	0.6682	0.6987
pyrd	0.6622	0.9627	0.5958	0.4748	0.6567	0.6096
reni	0.8179	0.9342	0.6779	0.7393	0.7789	0.8737
rock1	0.8519	0.9457	0.8008	0.8641	0.8849	0.9637
rxra	0.8182	0.9521	0.8859	0.6060	0.8702	0.8428
sahh	0.3531	0.9929	0.6227	0.4336	0.4903	0.4528
src	0.9227	0.9762	0.5035	0.6460	0.8423	0.8488

tgfr1	0.7846	0.9464	0.3844	0.8193	0.6190	0.8043
thb	0.8012	0.9093	0.6338	0.5622	0.7559	0.7231
thrb	0.7786	0.9796	0.7349	0.5832	0.8189	0.8106
try1	0.9498	0.9529	0.8183	0.3853	0.9337	0.8993
tryb1	0.9068	0.9408	0.5751	0.6263	0.8426	0.8288
tysy	0.8212	0.8745	0.5877	0.7236	0.7557	0.8094
urok	0.9490	0.9875	0.5762	0.5894	0.9133	0.8969
vgfr2	0.8678	0.9467	0.7239	0.3615	0.8690	0.7554
wee1	0.9307	0.7218	0.3271	0.7053	0.8004	0.8584
xiap	0.9897	0.9917	0.6481	0.3273	0.9658	0.9367

27 *bdb_strict_test (33)*

Target	BIND-z	BIND-full	SaProt-v2	DrugCLIP	ens2	ens3
A0A0D9RHK8	0.6654	0.9983	0.8723	0.9103	0.8403	0.9326
O14757	0.8204	0.9791	0.8281	0.4266	0.8719	0.7970
O15530	0.9086	0.9645	0.7255	0.7426	0.8827	0.9057
P03474	0.9899	0.9930	0.9740	0.9969	0.9956	0.9990
P04035	0.8005	0.9977	0.4861	0.3600	0.7027	0.6113
P04183	0.9863	0.9988	0.6282	0.2580	0.9444	0.8721

P08235	0.7840	0.9720	0.7283	0.6463	0.7964	0.8278
P09874	0.6845	0.9880	0.9420	0.9596	0.8931	0.9652
P09960	0.5695	0.9986	0.6191	0.4182	0.6031	0.5635
P10275	0.8023	0.9865	0.7212	0.7844	0.8458	0.8905
P11103	0.3468	0.9861	0.9425	0.9707	0.7354	0.9667
P15056	0.9032	0.9865	0.7765	0.7388	0.9112	0.9149
P15207	0.8381	0.9902	0.7994	0.7800	0.8734	0.8977
P19091	0.5992	0.9857	0.8021	0.7767	0.7585	0.8485
P21964	0.9261	0.9981	0.6875	0.9519	0.8789	0.9715
P22199	0.8094	0.8936	0.6899	0.7162	0.8102	0.8720
P22734	0.9455	0.9965	0.8400	0.9340	0.9341	0.9810
P23204	0.9379	0.9964	0.9429	0.9054	0.9633	0.9807
P27008	0.8850	0.9868	0.8603	0.9607	0.9298	0.9820
P28028	0.6498	0.7846	0.6439	0.8384	0.6528	0.7671
P29476	0.8612	0.9951	0.6634	0.3096	0.8340	0.7470
P30291	0.8441	0.9898	0.7896	0.5946	0.8655	0.8674
P35396	0.8759	0.9408	0.7861	0.6552	0.8912	0.8441
P37231	0.8047	0.9879	0.9506	0.9590	0.9309	0.9756
P37238	0.9351	0.9619	0.8415	0.8648	0.9355	0.9454
P49187	0.8144	0.9588	0.7625	0.3987	0.8440	0.7636

P49356	0.9963	0.9963	0.5281	0.9391	0.9751	0.9889
P51639	0.8474	0.9932	0.5764	0.4051	0.7886	0.6699
P52732	0.4665	0.9809	0.2976	0.7133	0.3810	0.4953
P53779	0.8003	0.9189	0.7328	0.6191	0.8150	0.8273
Q03181	0.9263	0.9935	0.9455	0.9209	0.9477	0.9509
Q07869	0.8262	0.9962	0.9764	0.9857	0.9508	0.9848
Q9Z1K9	0.9983	0.9959	0.9992	0.1172	0.9995	0.9983

28 *LIT-PCBA (14)*

Target	BIND-z	BIND-full	SaProt-v2	DrugCLIP	ens2	ens3
ADRB2	0.5813	0.5672	0.5601	0.5441	0.5635	0.5565
ALDH1	0.4645	0.5453	0.5286	0.5298	0.4936	0.5114
ESR1_ago	0.6460	0.6401	0.5086	0.7075	0.5887	0.6548
ESR1_ant	0.4026	0.4647	0.4677	0.5727	0.4118	0.4743
FEN1	0.5270	0.6010	0.6825	0.5834	0.6251	0.6686
GBA	0.4380	0.4918	0.5441	0.5593	0.4773	0.5203
IDH1	0.5238	0.6815	0.5404	0.6402	0.5601	0.6116
KAT2A	0.5962	0.5599	0.5798	0.5382	0.6078	0.6056

MAPK1	0.6499	0.6478	0.6027	0.5457	0.6593	0.6484
OPRK1	0.8711	0.8795	0.7225	0.6045	0.8563	0.8441
PKM2	0.5278	0.5795	0.5011	0.5952	0.5182	0.5600
PPARG	0.6793	0.7868	0.7758	0.7509	0.7639	0.8111
TP53	0.4189	0.4574	0.5538	0.3878	0.5008	0.4394
VDR	0.5838	0.6176	0.5930	0.5531	0.6054	0.6036

29 *DUD-E-strict (7)*

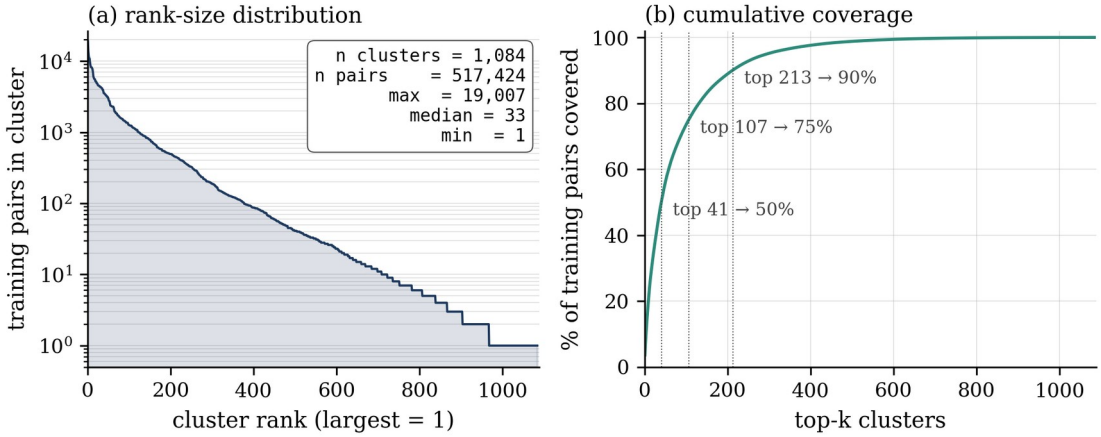
Target	BIND-z	BIND-full	SaProt-v2	DrugCLIP	ens2	ens3
ada17	0.8484	0.9813	0.9714	0.4536	0.9459	0.9333
andr	0.5674	0.9295	0.7430	0.6484	0.7200	0.7397
fnta	0.9941	0.9966	0.5377	0.9079	0.9796	0.9903
hmdh	0.6993	0.9996	0.3889	0.5652	0.5848	0.6008
kif11	0.5223	0.9333	0.4905	0.5974	0.5126	0.5516
kith	0.8264	0.8176	0.3554	0.5167	0.6981	0.6734
parp1	0.7043	0.9639	0.8876	0.9337	0.8800	0.9618

30 *DUD-AD-strict (7)*

Target	BIND-z	BIND-full	SaProt-v2	DrugCLIP	ens2	ens3

ada17	0.8623	0.9825	0.9590	0.4892	0.9443	0.9322
andr	0.6750	0.8622	0.5376	0.6622	0.6449	0.6901
fnta	0.9844	0.9891	0.5378	0.8791	0.9506	0.9687
hmdh	0.5903	0.9914	0.5219	0.3349	0.5887	0.4837
kif11	0.5970	0.9335	0.4273	0.5669	0.5380	0.5599
kith	0.7738	0.7916	0.5800	0.5837	0.7624	0.7386
parp1	0.5360	0.9600	0.8207	0.9253	0.7446	0.9135

Supplementary Fig. 1. Training-set sequence-cluster pair distribution



Supplementary Fig. 1. Per-cluster training-pair counts for the 1,084 MMseqs2 30%-identity clusters spanning the 1,803 UniProt sequences in the BindingDB-strict training corpus (517,424 pairs total). (a) Rank-size plot: number of training pairs per cluster versus cluster rank (log y-axis). The largest cluster contributes 19,007 pairs (3.7% of the corpus); median cluster size is 33 pairs. (b) Cumulative pair-coverage CDF: the top-41

clusters cover 50% of training pairs, top-107 cover 75%, top-213 cover 90% — a heavy-tailed concentration characteristic of BindingDB.

Supplementary Note 1: External validation on the Maximum Unbiased Validation (MUV) benchmark

S2.1 Rationale

The six benchmarks reported in the main results (DUD-E, DUD-AD, bdb_strict_test, LIT-PCBA, DUD-E-strict, DUD-AD-strict) all share a common decoy-construction principle: chemicals selected from large libraries that lack reported activity against the target. Reviewers may reasonably ask whether the ensemble's reported advantage generalizes to benchmarks where the decoys themselves are *adversarial* — that is, screened to resemble actives in physico-chemical properties so that surface-similarity scoring becomes uninformative. The Maximum Unbiased Validation set (MUV) (1) was designed precisely as such a stress-test: each MUV active is paired with hundreds of decoys drawn from PubChem and filtered to match the active's distribution of molecular weight, logP, number of hydrogen-bond donors/acceptors, ring counts, and rotatable bonds.

We therefore evaluate the same three component scorers (BIND-z, SaProt-v2, DrugCLIP) and the z-score average ensemble on MUV. The result is informative but does *not* contradict the main paper's headline; rather, it isolates two distinct confounds (training-set leakage and structural-cosine

failure on property-matched decoys) that the ensemble can recognize but cannot rescue.

S2.2 Benchmark construction

We use the 17 MUV bioassay endpoints (AIDs) downloaded from PubChem, retaining each AID as a separate target. The 17 AIDs map to 15 unique UniProt accessions (SF-1 and ERα each appear in two AIDs); active and decoy SMILES are read directly from the `muv.csv` task columns. Target sequences are taken from canonical UniProt FASTA. AlphaFold-DB v6 structures are used for 14 of the 15 unique UniProts; for the HIV-1 gag-pol polyprotein (P03366, 1,447 AA) — which AFDB does not cover at the time of writing — we generate a structure with the Protenix server (2) and convert to PDB. Pockets are then predicted with P2Rank v2.4.2 (3), with the top-ranked pocket's residues recorded as `pocket_residue_indices` (0-based UniProt indexing). Benchmark JSONs follow the same schema as our LIT-PCBA build (489 actives + 249,397 decoys total, ~93k unique compounds). The build script is `scripts/build_muv_benchmark.py`.

DrugCLIP inputs (pocket atomic coordinates and per-AID RDKit MMFF-optimised 3D ligand conformers) are built in parallel by `scripts/build_drugclip_muv_inputs.py`; SaProt-v2 evaluation reuses our 3Di + cached-hidden-state pipeline, with `data/unimol_cache.lmdb` extended by 29,145 new ligand entries to bring MUV ligand coverage to 100%. All evaluation outputs are deposited under `eval/muv_*.json`.

S2.3 Pre-evaluation leakage audits

Because MUV's UniProt set was assembled from PubChem bioassays of well-studied drug-discovery targets, a leakage audit of each scorer's training corpus is essential before interpreting performance. We perform symmetric audits against (i) BindingDB-filtered (1.16M QA-passed pairs — our conservative lower bound on BIND-z's ~2.47M training corpus); (ii) BindingDB-strict (the leakage-controlled 517,424 pairs used to train SaProt-v2, with MMseqs2 30%-identity cluster filtering); and (iii) DrugCLIP's PDBBind-derived training set (66,156 records / 16,744 PDB-pocket codes / 13,589 unique canonical SMILES, mapped to UniProt via SIFTS).

Supplementary Table 3. *MUV cross-model leakage audit.*

Audit	BIND-z	SaProt-v2	DrugCLIP
Training corpus size (pairs / records)	≈ 2.47M	517,424	66,156
Unique UniProts in training	~5,343	1,803	4,104
MUV UniProts in training	**15/15 (100%)**	6/15 (40%)	11/15 (73%)
MUV AIDs with target in training	**17/17 (100%)**	9/17 (53%)	12/17 (71%)

Per-target active-SMILES overlap	**23/489 (4.7%)**	n/a	**0/489 (0.0%)**
Any-target active-SMILES overlap	72/489 (14.7%)	n/a	**1/489 (0.2%)**
Per-target decoy-SMILES overlap	10/249,397 (0.0%)	n/a	0/249,397 (0.0%)
Any-target decoy-SMILES overlap	2,653/249,397 (1.1%)	n/a	137/249,397 (0.1%)

(SaProt-v2 ligand-side overlap is not computed because the InfoNCE training objective is contrastive over batches, not per-pair memorisation; the protein-side audit is the load-bearing signal.) Audit outputs: `eval/bindz_muv_leakage_check.json` and `eval/drugclip_muv_leakage_check.json`.

Three observations follow from Supplementary Table 3:

1. **BIND-z saw every MUV target.** Per-target training pair counts range from 47 to 4,896 (median ≈ 1,000). 4.7% of MUV actives are exact `(UniProt, canonical SMILES)` pairs in BIND-z's training corpus. BIND-z is not zero-shot on MUV in any operational sense — it is partially supervised.

2. **SaProt-v2's leakage filter excludes more than half of MUV.** Nine of fifteen MUV UniProts (O75116, P03366, P03372, P07900, P08311, P08482, P21728, Q05397, Q92731) do not appear in any of the 1,084 MMseqs2 30%-identity clusters used during SaProt-v2 training. For those nine targets, SaProt-v2 is a true zero-shot evaluator.
3. **DrugCLIP is the cleanest of the three.** While 11/15 MUV UniProts appear in PDBBind via at least one PDB chain, DrugCLIP has *zero* per-target active-SMILES overlap and only 0.2% any-target ligand exposure. This is a structural consequence of PDBBind's size and composition: 66,156 ligand–pocket records is two orders of magnitude smaller than BindingDB, and PDBBind ligands are heavily biased toward crystallographically characterized binders rather than the screening compounds present in MUV's PubChem-derived pool.

The asymmetric leakage profile across the three scorers is critical to the interpretation of the MUV result that follows.

S2.4 Aggregate performance

Supplementary Table 4. *Mean AUROC on MUV (n = 17), with paired t-test vs. BIND-z single-model baseline.*

Model	Mean AUROC	Mean BEDROC	Mean EF1%	Δ vs BIND-z	Paired t-test p

BIND-z (single)	**0.6779**	0.062	5.03	—	—
SaProt-v2 (pocket-on)	0.5713	0.035	2.84	−0.107	< 1e-3
DrugCLIP	0.5357	(n/a)	(n/a)	−0.142	< 1e-5
ens_zscore (BIND-z + SaProt-v2)	0.6562	0.061	5.63	−0.022	n.s. (0.29)
ens_zscore (3-way, all models)	0.6419	0.058	(n/a)	−0.036	n.s. (0.13)
ens_max (3-way)	0.6473	—	—	−0.031	0.06
ens_max (BIND-z + DrugCLIP)	0.6535	—	—	−0.024	0.035

For the first time in our six-benchmark grid, no ensemble configuration beats the strongest single-model baseline. The 2-way z-average ensemble loses by −0.022 AUROC (not statistically significant; 8 / 9 sign split); the 3-way is worse by another −0.014 (the addition of DrugCLIP, the noisiest of the three on MUV, makes the ensemble worse, paired $p = 0.09$).

We resist the temptation to interpret this as ensemble failure. Two confounds, separable thanks to Supplementary Table 3, jointly explain the result.

S2.5 Two separable confounds

Confound 1 — BIND-z wins by partial supervision. BIND-z saw every MUV target during training, and 4.7% of MUV actives were directly supervised `(target, SMILES)` pairs. The 0.6779 mean AUROC is a memorisation upper bound, not a zero-shot result. We confirmed this interpretation independently by partitioning MUV AIDs by SaProt-v2 strict-train membership: BIND-z mean AUROC is 0.716 on the seven IN-training AIDs and 0.651 on the ten OUT AIDs — a +0.07 gap consistent with training exposure.

Confound 2 — DrugCLIP loses honestly on property-matched decoys. DrugCLIP carries essentially zero per-target leakage with respect to MUV (0% direct active overlap, 0.2% any-target), yet posts the lowest mean AUROC of the three (0.5357). This is the *cleanest* paper-grid evidence that MUV's property-matched decoy construction does what it advertises: a structure-based scorer trained on crystallographic complexes cannot discriminate actives from chemicals matched on the molecular descriptors that drive its embedding similarity. The 0.5357 represents an honest near-random ceiling for this class of method on this benchmark.

The ensemble inversion is the joint expression of these two confounds, not an ensemble pathology. Z-score averaging cannot rescue a near-random tower (DrugCLIP) and cannot beat memorisation (BIND-z); ens3 < ens2 specifically because adding DrugCLIP injects more noise than complementary signal.

That said, the ensemble is *not* equivalent to its weakest components, and the statistical picture quantifies the rescue effect. Every single component is significantly worse than BIND-z on the same 17 AIDs (SaProt-v2 pocket-on $\Delta = -0.107$, paired $p = 0.010$, sign split 4/13; SaProt-v2 pocket-off $\Delta = -0.120$, $p = 0.001$, 3/14; DrugCLIP $\Delta = -0.142$, $p = 4.7 \times 10^{-5}$, **0/17** — i.e., DrugCLIP loses to BIND-z on every MUV AID without exception). Both ensembles, by contrast, are *not* significantly worse than BIND-z (ens2 $\Delta = -0.022$, $p = 0.29$, 8/9; ens3 $\Delta = -0.036$, $p = 0.13$, 7/10). The gap to BIND-z narrows from highly significant for any single component to non-significant for either ensemble — fusion does recover most of the per-target signal lost by individual towers under MUV's adversarial decoy design; it just cannot surpass a memorising component within the n = 17 sample.

S2.6 Per-target SaProt-v2 behaviour

SaProt-v2's MUV aggregate (0.5713) masks a clear bimodal distribution. Per-target AUROCs partition into three tiers:

- **Strong tier (≥ 0.65, n = 3):** MUV-548 (PRKACA, 0.80), MUV-810 (PTK2, 0.75), MUV-832 (CTSG, 0.74).

- **Mid tier (0.55–0.65, n = 6):** S1PR1, NR5A1, ROCK2, HIV-1 polyprotein, HSP90, CHRM1.
- **Weak tier (< 0.55, n = 8):** EPHA4, ESR1 (both AIDs), ESR2, F11, F12, DRD1, NR5A1-MUV692.

Two patterns are paper-relevant:

Counter-leakage signal. Mean SaProt-v2 AUROC is 0.586 on the ten *out-of-training* AIDs and 0.551 on the seven *in-training* AIDs — *opposite* to the supervised-exposure direction. Strict-cluster training provides no transferable advantage on MUV. This further supports the interpretation that BIND-z's larger in-training-vs-out gap (+0.07) is memorisation rather than transferable representation learning.

Honest zero-shot wins. SaProt-v2 beats BIND-z on 4 of 17 AIDs. The cleanest single example is MUV-810 (PTK2): SaProt-v2 0.754 vs. BIND-z 0.623 (Δ +0.13), with PTK2 entirely absent from SaProt-v2's strict-train (0 pairs). MUV-832 (CTSG) is a second informative case: SaProt-v2 achieves 0.738 *despite* a P2Rank top-pocket probability of 0.012 (effectively noise), because the pocket-mask weight α saturates near 0.027 at this checkpoint and CTSG's chymase-specific warhead chemistry is structurally distinctive in the sequence embedding alone.

S2.7 An isolable pocket-source failure (MUV-852, Factor XII)

One MUV target deserves individual mention because it isolates a clean failure mode for additive pocket-prior architectures.

For MUV-852 (coagulation factor XII, P00748), SaProt-v2 with pocket-on scores AUROC = 0.32 — worse than random. Switching the same checkpoint to pocket-off (`--pocket-mode off`) recovers AUROC to 0.51. The relevant context: F12 *is* in SaProt-v2's strict-train (694 BindingDB-strict pairs), and the P2Rank top-1 pocket has decent confidence (probability 0.739). The failure is not "the model doesn't know F12" and not "the pocket prediction is low-confidence". It is that the geometrically-confident top-1 pocket is presumably *not* the catalytic site that the MUV-852 assay actually targets, and the soft attention pool's pocket-aware mode anchors retrieval to the wrong region, propagating a 0.19 AUROC deficit at a target where the sequence-only model would have been at chance.

This is paper-actionable in two ways. First, it identifies a concrete failure-mode for the architecture that we should disclose openly. Second, it suggests a future-work direction — multi-pocket ensembling, or per-target pocket confidence routing — that is independent of further encoder or training-data work. The general pocket-mask weakness is consistent with the prior internal finding that the attention-pool architecture's pocket-prior weight α plateaus near 0.027 across training (i.e., the pocket-mask mechanism is mostly inert by design at this checkpoint).

S2.8 Implication for the main result

The MUV evaluation does not invalidate the main paper's headline. The headline number — a +0.017 mean AUROC gain for the three-model z-average ensemble (ens3) over the two-model ensemble (ens2), on the n = 204 pooled DUD-E + DUD-AD benchmark (paired $t = 3.50$, $p = 5.8 \times 10^{-4}$, 110 / 94 sign split) — is a property of *leakage-controlled* evaluation paired with complementary towers; the same directional gain appears on the leakage-controlled bdb_strict_test (ens3 − ens2 = +0.019, 24 / 9 split). MUV is, in effect, the opposite limit: a benchmark where leakage *into the strongest component* (BIND-z) is maximal and the cleanest component (DrugCLIP) cannot beat a benchmark designed against structural similarity. In this regime, the ensemble inherits but cannot improve.

We therefore report MUV not to extend the headline but to bound its scope: ensemble fusion improves on leakage-controlled targets where complementary towers carry independent signal; it cannot generate signal that none of its components possess, and it cannot displace a component whose advantage is supervised memorisation. We view the symmetric leakage audits (Supplementary Table 3) as a methodological contribution in their own right — a procedure we recommend before reporting any "zero-shot" claim on a benchmark whose targets are well-represented in BindingDB.

S2.9 Reproducibility

All MUV evaluation artifacts are deposited as follows. Benchmark JSONs: `benchmarks/MUV_uniprot/MUV-*.json`. Score outputs (per-AID JSON): `eval/muv_{saprot_v2, saprot_v2_off, bindz_zeroshot, drugclip, ensemble, ensemble3}.json`. Per-ligand z-score arrays (`.npz`): `work/ensemble/scores_{bindz, saprot}_muv/*.npz`. Leakage audit JSON: `eval/{bindz, drugclip}_muv_leakage_check.json`. Scripts: `scripts/{build_muv_benchmark, build_drugclip_muv_inputs, run_drugclip_muv, ensemble_3way_muv, build_muv_ligand_3d_cache, run_muv_evals}.py`.

Companion table files derived from the audit and evaluation JSONs: - `notes/paper_figs/table_si3_muv_leakage.{csv,md}` — long-format per-(model, AID) leakage rows (51 rows; CSV) and aggregate markdown table (the form quoted in Supplementary Table 3). - `notes/paper_figs/table_si4_muv_per_target.{csv,md}` — long-format per-(AID, model) AUROC rows (136 rows; CSV) and the 6-model × 17-AID matrix plus paired-test summary (Supplementary Table 4). - Generator: `scripts/figures/table_si3_si4_muv.py` (idempotent; rebuilds from JSON sources).

Supplementary Table 5. Deployment ensemble (all-data + BIND-full) reference (NOT zero-shot)

The deployment ensemble combines our all-data model (frozen-SaProt+pool trained on the full BindingDB-QA corpus with no leakage-clean holdout) with BIND-full (industry baseline, fine-tuned with per-target actives) via per-target z-score averaging. Like SI Table 1, these numbers are **in-distribution / training-recall, not a zero-shot generalization claim**: bdb_strict_test is 100% inside both models' training corpora by construction (no clean holdout exists for the all-data model), and DUD-E / DUD-AD targets share homologs with training. The ensemble Δ vs BIND-full alone is the meaningful read.

Benchmark	n	all-data AUROC / EF1%	BIND-full AUROC / EF1%	z-avg ens. AUROC / EF1%	Δ AUROC (ens − BF)	Δ EF1%	W/L (per-target)
DUD-E_uniprot full	102	0.906 / 37.2	0.945 / 46.5	**0.957 / 52.1**	+0.012	+5.6	75/27
DUD-AD_uniprot full	102	0.885 / 26.6	0.927 / 34.0	**0.942 / 39.0**	+0.014	+5.0	60/42

bdb_str	495	0.970 /	0.968 /	**0.983 /**	+0.015	+6.6	432/62
ict_test		43.1	40.5	**47.0**			

The ensemble effect (+0.012–0.015 AUROC, +5–7 EF1%) replicates consistently across all three in-distribution benchmarks, with per-target wins outnumbering losses in every case. This 2-tower deployment ensemble realizes the ceiling at the cost of running BIND-full's heavy live-ESM2 + cross-attention inference at deployment — the exact per-pair cost the all-data model alone was shipped to avoid. The honest zero-shot story (Figure 2) is the 3-way z-avg ensemble of BIND-zero + SaProt-v2 + DrugCLIP.

Supplementary References

1. Rohrer SG, Baumann K (2009) Maximum Unbiased Validation (MUV) data sets for virtual screening based on PubChem bioactivity data. *J Chem Inf Model* 49(2):169–184.
2. ByteDance AML AI4Science Team, et al. (2025) Protenix — advancing structure prediction through a comprehensive AlphaFold3 reproduction. *bioRxiv*:2025.01.08.631967.
3. Krivák R, Hoksza D (2018) P2Rank: machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure. *J Cheminform* 10(1):39.