Experimental errors in QSAR modeling sets: what we can do and what we cannot do

Linlin Zhao[†], Wenyi Wang[†], Alexander Sedykh[*,‡], and Hao Zhu[*, †, §]

[†] The Rutgers Center for Computational and Integrative Biology, Camden, New Jersey 08102, USA;

[‡] Sciome LLC, Durham, North Carolina 27709, USA

[§] Department of Chemistry, Rutgers University, Camden, New Jersey 08102, USA

*Corresponding author:

Hao Zhu:

315 Penn St.

Rutgers University

Camden, NJ 08102

Telephone: (856) 225-6781

Email: hao.zhu99@rutgers.edu

Alexander Sedykh:

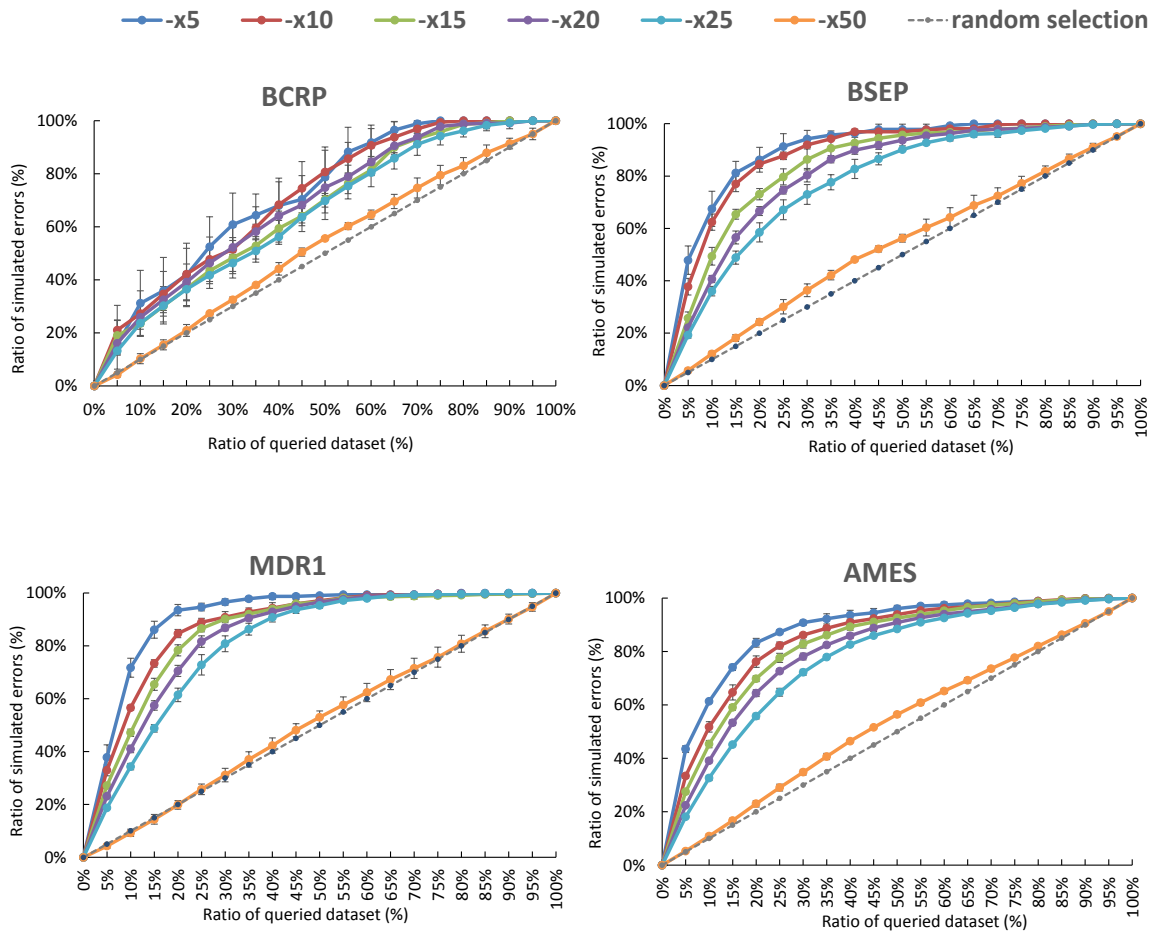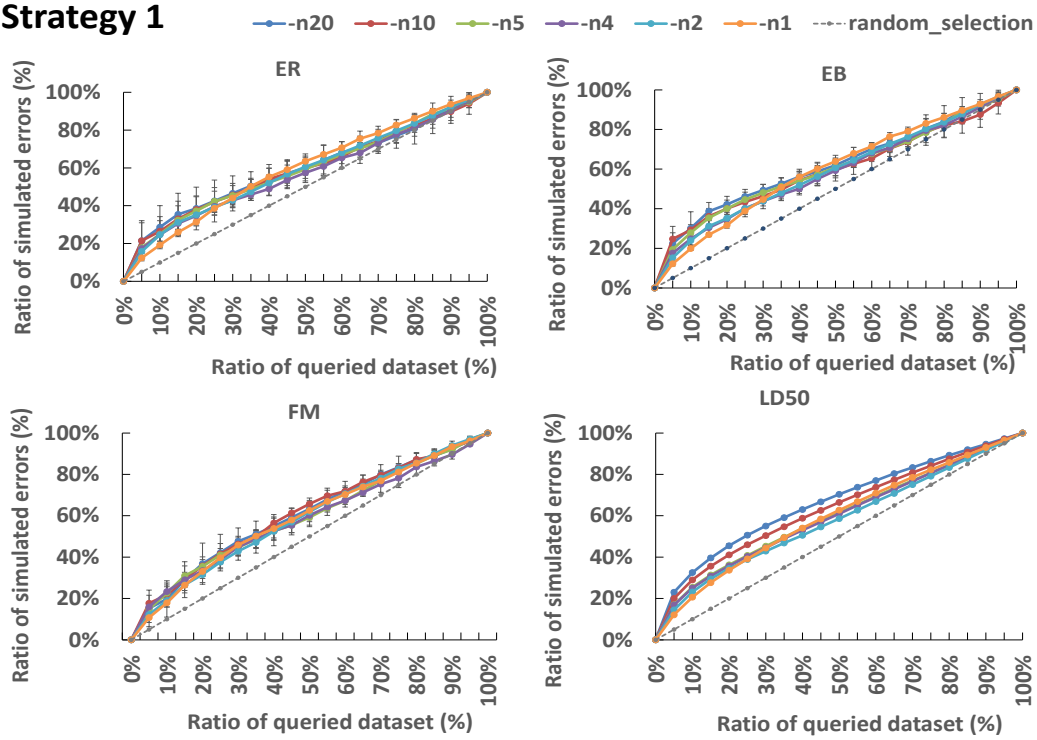2 Davis Drive

Sciome LLC

Research Triangle Park, Durham, North Carolina 27709

Figure S1. AUC plots for categorical datasets.
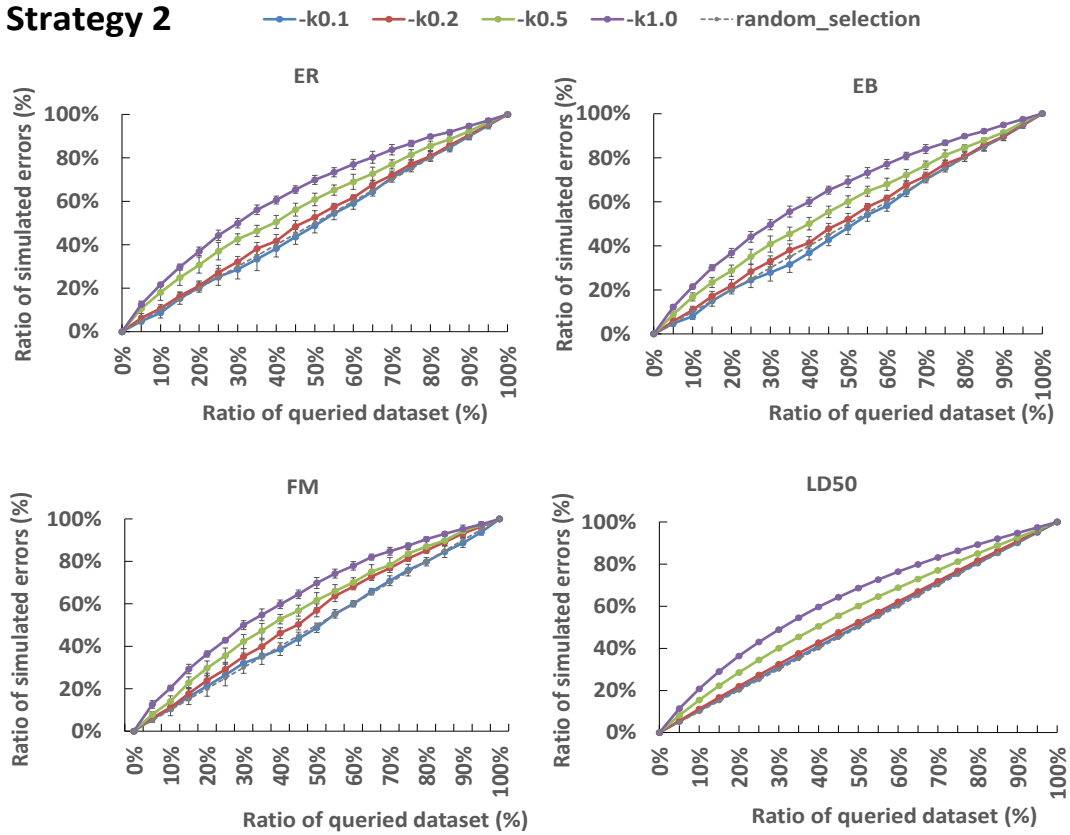
Strategy 1

Strategy 2

Figure S2. AUC plots for continuous datasets.

Table S1 Five-fold cross-validation results for categorical datasets

Model			BCRP			BSEP			MDR1			AMES		
			Sensitivity	Specificity	CCR	Sensitivity	Specificity	CCR	Sensitivity	Specificity	CCR	Sensitivity	Specificity	CCR
Control	Individual models	MOE_RF	0.750	0.823	0.787	0.831	0.883	0.857	0.903	0.891	0.897	0.747	0.883	0.815
		MOE_SVM	0.459	0.801	0.630	0.791	0.857	0.824	0.853	0.858	0.855	0.715	0.822	0.768
		Dragon_RF	0.743	0.829	0.786	0.803	0.871	0.837	0.905	0.891	0.898	0.737	0.879	0.808
		Dragon_SVM	0.486	0.845	0.666	0.791	0.840	0.816	0.913	0.838	0.875	0.541	0.886	0.713
	Consensus model		0.649	0.851	0.750	0.827	0.871	0.849	0.916	0.871	0.894	0.727	0.893	0.810
-x5	Individual models	MOE_RF	0.721	0.802	0.762	0.798	0.858	0.828	0.852	0.848	0.850	0.697	0.858	0.777
		MOE_SVM	0.367	0.846	0.607	0.710	0.852	0.781	0.812	0.822	0.817	0.685	0.800	0.742
		Dragon_RF	0.680	0.791	0.736	0.742	0.849	0.796	0.851	0.830	0.841	0.704	0.843	0.773
		Dragon_SVM	0.463	0.830	0.646	0.714	0.818	0.766	0.876	0.779	0.828	0.492	0.884	0.688
	Consensus model		0.585	0.824	0.705	0.774	0.847	0.810	0.871	0.822	0.846	0.684	0.874	0.779
-x10	Individual models	MOE_RF	0.664	0.767	0.716	0.731	0.849	0.790	0.797	0.823	0.810	0.668	0.834	0.751
		MOE_SVM	0.275	0.878	0.576	0.648	0.849	0.749	0.748	0.804	0.776	0.661	0.780	0.721
		Dragon_RF	0.624	0.733	0.679	0.711	0.815	0.763	0.796	0.810	0.803	0.652	0.828	0.740
		Dragon_SVM	0.389	0.844	0.617	0.648	0.803	0.726	0.816	0.766	0.791	0.400	0.890	0.645
	Consensus model		0.463	0.839	0.651	0.700	0.829	0.764	0.819	0.800	0.810	0.635	0.849	0.742
-x15	Individual models	MOE_RF	0.634	0.783	0.709	0.675	0.817	0.746	0.772	0.792	0.782	0.629	0.806	0.718
		MOE_SVM	0.214	0.940	0.577	0.580	0.862	0.721	0.758	0.775	0.766	0.630	0.760	0.695
		Dragon_RF	0.614	0.788	0.701	0.631	0.794	0.713	0.771	0.758	0.764	0.613	0.795	0.704
		Dragon_SVM	0.338	0.875	0.606	0.608	0.794	0.701	0.790	0.740	0.765	0.410	0.861	0.635
	Consensus model		0.434	0.902	0.668	0.643	0.831	0.737	0.795	0.755	0.775	0.602	0.832	0.717
-x20	Individual models	MOE_RF	0.662	0.729	0.696	0.612	0.769	0.691	0.704	0.779	0.741	0.587	0.796	0.692
		MOE_SVM	0.257	0.895	0.576	0.516	0.855	0.685	0.733	0.740	0.736	0.611	0.743	0.677
		Dragon_RF	0.662	0.740	0.701	0.605	0.763	0.684	0.686	0.761	0.724	0.566	0.781	0.674
		Dragon_SVM	0.405	0.801	0.603	0.469	0.824	0.646	0.731	0.718	0.725	0.358	0.882	0.620
	Consensus model		0.493	0.823	0.658	0.574	0.821	0.697	0.741	0.741	0.741	0.552	0.823	0.687
-x25	Individual models	MOE_RF	0.578	0.736	0.657	0.585	0.755	0.670	0.668	0.776	0.722	0.536	0.765	0.650
		MOE_SVM	0.238	0.912	0.575	0.316	0.860	0.588	0.646	0.726	0.686	0.581	0.722	0.651
		Dragon_RF	0.626	0.670	0.648	0.605	0.729	0.667	0.652	0.745	0.699	0.508	0.758	0.633
		Dragon_SVM	0.333	0.846	0.590	0.407	0.823	0.615	0.644	0.721	0.683	0.218	0.909	0.564

	Consensus model		0.361	0.857	0.609	0.486	0.818	0.652	0.684	0.754	0.719	0.480	0.820	0.650
-x50	**Individual models**	**MOE_RF**	0.442	0.566	0.504	0.246	0.750	0.498	0.459	0.568	0.513	0.324	0.711	0.517
		MOE_SVM	0.130	0.909	0.519	0.004	1.000	0.502	0.050	0.950	0.500	0.000	1.000	0.500
		Dragon_RF	0.409	0.560	0.485	0.213	0.728	0.470	0.456	0.585	0.520	0.305	0.703	0.504
		Dragon_SVM	0.136	0.869	0.502	0.000	1.000	0.500	0.000	1.000	0.500	0.000	1.000	0.500
	Consensus model		0.201	0.851	0.526	0.004	0.994	0.499	0.034	0.964	0.499	0.001	1.000	0.500

Table S2 Five-fold cross-validation results for continuous datasets

Model			ER MAE	ER R²	FM MAE	FM R²	EB MAE	EB R²	LD50 MAE	LD50 R²
Control	Individual models	MOE_RF	0.668	0.833	0.614	0.774	0.575	0.776	0.457	0.762
		MOE_SVM	0.823	0.732	0.735	0.686	0.841	0.652	0.581	0.578
		Dragon_RF	0.730	0.803	0.643	0.753	0.595	0.764	0.459	0.757
		Dragon_SVM	1.066	0.643	0.773	0.693	0.706	0.718	0.555	0.615
	Consensus model		0.773	0.807	0.659	0.766	0.637	0.776	0.492	0.730
-n20	Individual models	MOE_RF	0.676	0.827	0.632	0.757	0.583	0.768	0.463	0.755
		MOE_SVM	0.826	0.736	0.746	0.683	0.843	0.625	0.583	0.577
		Dragon_RF	0.747	0.796	0.653	0.741	0.601	0.750	0.466	0.752
		Dragon_SVM	1.072	0.638	0.775	0.687	0.713	0.696	0.557	0.617
	Consensus model		0.777	0.804	0.665	0.758	0.641	0.762	0.495	0.727
-n10	Individual models	MOE_RF	0.693	0.824	0.637	0.757	0.596	0.751	0.474	0.738
		MOE_SVM	0.835	0.731	0.760	0.672	0.842	0.617	0.583	0.567
		Dragon_RF	0.757	0.792	0.663	0.736	0.614	0.741	0.476	0.738
		Dragon_SVM	1.073	0.630	0.786	0.673	0.708	0.682	0.560	0.603
	Consensus model		0.786	0.801	0.672	0.751	0.640	0.751	0.500	0.712
-n5	Individual models	MOE_RF	0.757	0.766	0.701	0.692	0.654	0.685	0.513	0.678
		MOE_SVM	0.894	0.667	0.792	0.614	0.840	0.585	0.593	0.530
		Dragon_RF	0.820	0.742	0.721	0.681	0.656	0.683	0.511	0.679
		Dragon_SVM	1.072	0.595	0.818	0.615	0.737	0.637	0.572	0.566
	Consensus model		0.837	0.746	0.718	0.691	0.666	0.695	0.523	0.660
-n4	Individual models	MOE_RF	0.807	0.726	0.760	0.651	0.702	0.647	0.533	0.653
		MOE_SVM	0.924	0.629	0.807	0.596	0.857	0.538	0.602	0.515
		Dragon_RF	0.863	0.693	0.779	0.624	0.720	0.621	0.537	0.647
		Dragon_SVM	1.090	0.546	0.824	0.593	0.764	0.571	0.586	0.538
	Consensus model		0.870	0.703	0.759	0.653	0.700	0.645	0.541	0.634
-n2	Individual models	MOE_RF	0.990	0.566	0.910	0.462	0.804	0.414	0.647	0.440
		MOE_SVM	1.047	0.487	0.949	0.388	0.861	0.359	0.662	0.346
		Dragon_RF	1.027	0.540	0.925	0.440	0.825	0.379	0.648	0.438
		Dragon_SVM	1.141	0.448	0.941	0.429	0.825	0.363	0.656	0.354
	Consensus model		1.010	0.565	0.897	0.470	0.794	0.417	0.633	0.440
-n1	Individual models	MOE_RF	1.428	0.022	1.094	0.017	0.997	-0.008	0.772	0.008
		MOE_SVM	1.377	-0.013	1.068	0.020	0.945	-0.006	0.734	0.000
		Dragon_RF	1.441	-0.049	1.093	0.036	0.996	-0.024	0.774	0.016
		Dragon_SVM	1.348	0.026	1.062	-0.047	0.953	-0.065	0.745	-0.003
	Consensus model		1.380	-0.010	1.064	0.023	0.960	-0.024	0.744	0.010
-k0.1	Individual models	MOE_RF	0.673	0.831	0.623	0.771	0.586	0.771	0.465	0.756
		MOE_SVM	0.814	0.740	0.746	0.684	0.844	0.657	0.586	0.574
		Dragon_RF	0.738	0.801	0.665	0.744	0.604	0.764	0.468	0.753
		Dragon_SVM	1.063	0.647	0.781	0.692	0.714	0.718	0.562	0.610
	Consensus model		0.779	0.809	0.671	0.762	0.646	0.774	0.499	0.726
-k0.2	Individual models	MOE_RF	0.715	0.824	0.669	0.751	0.629	0.753	0.488	0.746
		MOE_SVM	0.869	0.732	0.764	0.674	0.875	0.639	0.604	0.570
		Dragon_RF	0.772	0.800	0.686	0.736	0.630	0.754	0.492	0.739

		Dragon_SVM	1.128	0.641	0.809	0.684	0.752	0.702	0.582	0.600
	Consensus model		0.818	0.806	0.697	0.752	0.680	0.760	0.520	0.715
-k0.5	**Individual models**	**MOE_RF**	0.938	0.736	0.835	0.696	0.794	0.660	0.616	0.671
		MOE_SVM	1.017	0.663	0.907	0.629	0.969	0.576	0.706	0.528
		Dragon_RF	1.013	0.678	0.882	0.668	0.816	0.637	0.622	0.665
		Dragon_SVM	1.232	0.583	0.942	0.637	0.870	0.613	0.694	0.540
	Consensus model		0.996	0.717	0.858	0.697	0.824	0.661	0.638	0.651
-k1.0	**Individual models**	**MOE_RF**	1.413	0.563	1.406	0.498	1.164	0.485	0.915	0.507
		MOE_SVM	1.454	0.523	1.435	0.444	1.290	0.421	0.969	0.404
		Dragon_RF	1.441	0.544	1.421	0.484	1.165	0.483	0.918	0.501
		Dragon_SVM	1.599	0.448	1.486	0.455	1.201	0.478	0.972	0.397
	Consensus model		1.420	0.573	1.398	0.510	1.163	0.505	0.923	0.500

Table S3 External prediction results for categorical datasets

Model	BCRP			BSEP			MDR1			AMES		
	CCR without AD	CCR with AD	Coverage with AD	CCR without AD	CCR with AD	Coverage with AD	CCR without AD	CCR with AD	Coverage with AD	CCR without AD	CCR with AD	Coverage with AD
Control	0.76	0.76	0.68	0.89	0.85	0.69	0.92	0.92	0.68	0.82	0.85	0.72
Control-r5	0.74	0.71	0.68	0.87	0.83	0.69	0.92	0.96	0.31	0.82	0.86	0.72
Control-r10	0.74	0.74	0.65	0.85	0.82	0.67	0.91	0.96	0.31	0.82	0.87	0.71
Control-r15	0.67	0.65	0.65	0.82	0.80	0.68	0.90	0.96	0.31	0.82	0.86	0.81
Control-r20	0.65	0.66	0.65	0.80	0.79	0.65	0.90	0.96	0.30	0.80	0.85	0.80
-x5	0.78	0.80	0.68	0.88	0.85	0.69	0.92	0.92	0.68	0.83	0.86	0.72
-x5-r5	0.75	0.76	0.68	0.88	0.85	0.69	0.92	0.96	0.31	0.82	0.86	0.72
-x5-r10	0.70	0.71	0.65	0.87	0.86	0.67	0.91	0.96	0.31	0.82	0.87	0.71
-x5-r15	0.68	0.69	0.65	0.83	0.81	0.68	0.90	0.96	0.31	0.82	0.85	0.81
-x5-r20	0.68	0.68	0.65	0.82	0.80	0.65	0.90	0.97	0.30	0.80	0.85	0.80
-x10	0.79	0.80	0.68	0.89	0.86	0.69	0.89	0.89	0.68	0.80	0.84	0.72
-x10-r5	0.74	0.76	0.68	0.89	0.86	0.69	0.89	0.95	0.31	0.80	0.84	0.72
-x10-r10	0.65	0.62	0.65	0.88	0.86	0.67	0.88	0.94	0.31	0.80	0.84	0.71
-x10-r15	0.63	0.63	0.65	0.85	0.83	0.68	0.89	0.95	0.31	0.80	0.84	0.81
-x10-r20	0.63	0.64	0.65	0.85	0.84	0.65	0.88	0.96	0.30	0.80	0.84	0.80
-x15	0.63	0.63	0.68	0.83	0.83	0.69	0.88	0.87	0.68	0.81	0.84	0.72
-x15-r5	0.64	0.64	0.68	0.83	0.83	0.69	0.90	0.95	0.31	0.82	0.85	0.72
-x15-r10	0.59	0.59	0.65	0.84	0.83	0.67	0.90	0.95	0.31	0.80	0.85	0.71
-x15-r15	0.59	0.59	0.65	0.84	0.82	0.68	0.90	0.95	0.31	0.81	0.84	0.81
-x15-r20	0.59	0.61	0.65	0.84	0.80	0.65	0.89	0.96	0.30	0.80	0.84	0.80
-x20	0.77	0.76	0.68	0.89	0.86	0.69	0.87	0.88	0.68	0.80	0.83	0.72
-x20-r5	0.71	0.72	0.68	0.87	0.85	0.69	0.87	0.95	0.31	0.79	0.83	0.72
-x20-r10	0.68	0.68	0.65	0.90	0.87	0.67	0.87	0.94	0.31	0.79	0.83	0.71
-x20-r15	0.64	0.63	0.65	0.88	0.84	0.68	0.89	0.95	0.31	0.78	0.82	0.81
-x20-r20	0.61	0.62	0.65	0.85	0.82	0.65	0.88	0.96	0.30	0.79	0.84	0.80
-x25	0.71	0.72	0.68	0.82	0.81	0.69	0.85	0.84	0.68	0.79	0.82	0.72
-x25-r5	0.67	0.68	0.68	0.82	0.78	0.69	0.85	0.91	0.31	0.79	0.82	0.72
-x25-r10	0.59	0.59	0.65	0.81	0.78	0.67	0.87	0.93	0.31	0.79	0.83	0.71
-x25-r15	0.59	0.59	0.65	0.80	0.75	0.68	0.87	0.93	0.31	0.79	0.82	0.81
-x25-r20	0.61	0.60	0.65	0.82	0.80	0.65	0.87	0.94	0.30	0.78	0.82	0.80
-x50	0.41	0.43	0.68	0.45	0.47	0.69	0.40	0.41	0.68	0.50	0.50	0.72
-x50-r5	0.42	0.43	0.68	0.47	0.50	0.69	0.40	0.04	0.31	0.50	0.50	0.72
-x50-r10	0.43	0.45	0.65	0.49	0.49	0.67	0.42	0.53	0.31	0.50	0.50	0.71
-x50-r15	0.47	0.47	0.65	0.49	0.49	0.68	0.49	0.50	0.31	0.50	0.50	0.81
-x50-r20	0.51	0.52	0.65	0.49	0.49	0.65	0.49	0.50	0.30	0.50	0.50	0.80

Table S4 External prediction results for continuous datasets

Model	ER			FM			EB			LD50		
	MAE without AD	MAE with AD	Coverage with AD	MAE without AD	MAE with AD	Coverage with AD	MAE without AD	MAE with AD	Coverage with AD	MAE without AD	MAE with AD	Coverage with AD
Control	0.74	0.60	0.83	0.57	0.68	0.89	0.63	0.53	0.89	0.48	0.44	0.76
Control-r5	0.74	0.60	0.83	0.57	0.68	0.89	0.63	0.53	0.89	0.48	0.44	0.76
Control-r10	0.74	0.60	0.83	0.57	0.68	0.89	0.63	0.53	0.89	0.48	0.44	0.76
Control-r15	0.74	0.60	0.83	0.57	0.68	0.89	0.63	0.53	0.89	0.48	0.44	0.76
Control-r20	0.74	0.60	0.83	0.57	0.68	0.89	0.63	0.53	0.89	0.48	0.44	0.76
-n20	0.75	0.67	0.84	0.57	0.51	0.84	0.63	0.59	0.75	0.48	0.55	0.84
-n20-r5	0.77	0.67	0.84	0.70	0.63	0.84	0.63	0.69	0.75	0.49	0.58	0.84
-n20-r10	0.78	0.68	0.84	0.72	0.60	0.84	0.63	0.69	0.75	0.52	0.57	0.84
-n20-r15	0.80	0.65	0.84	0.72	0.59	0.84	0.63	0.69	0.75	0.54	0.56	0.84
-n20-r20	0.82	0.64	0.84	0.72	0.59	0.84	0.68	0.66	0.75	0.55	0.54	0.84
-n10	0.75	0.68	0.84	0.58	0.52	0.84	0.61	0.58	0.75	0.48	0.44	0.68
-n10-r5	0.77	0.69	0.84	0.63	0.67	0.84	0.62	0.69	0.75	0.51	0.61	0.68
-n10-r10	0.80	0.66	0.84	0.70	0.62	0.84	0.67	0.67	0.75	0.52	0.57	0.68
-n10-r15	0.80	0.66	0.84	0.72	0.61	0.84	0.68	0.66	0.75	0.53	0.56	0.68
-n10-r20	0.83	0.64	0.84	0.73	0.60	0.84	0.69	0.64	0.75	0.54	0.54	0.68
-n5	0.79	0.74	0.84	0.58	0.53	0.84	0.63	0.60	0.75	0.49	0.44	0.68
-n5-r5	0.81	0.68	0.84	0.65	0.67	0.84	0.65	0.66	0.75	0.50	0.59	0.68
-n5-r10	0.85	0.63	0.84	0.67	0.65	0.84	0.66	0.65	0.75	0.53	0.57	0.68
-n5-r15	0.85	0.63	0.84	0.67	0.65	0.84	0.67	0.64	0.75	0.54	0.55	0.68
-n5-r20	0.88	0.60	0.84	0.68	0.64	0.84	0.71	0.64	0.75	0.54	0.54	0.68
-n4	0.79	0.74	0.84	0.61	0.53	0.84	0.63	0.62	0.75	0.50	0.46	0.68
-n4-r5	0.81	0.69	0.84	0.64	0.67	0.84	0.69	0.66	0.75	0.51	0.60	0.68
-n4-r10	0.85	0.64	0.84	0.68	0.65	0.84	0.70	0.64	0.75	0.54	0.58	0.68
-n4-r15	0.87	0.61	0.84	0.71	0.65	0.84	0.72	0.62	0.75	0.55	0.56	0.68
-n4-r20	0.88	0.61	0.84	0.72	0.64	0.84	0.72	0.63	0.75	0.55	0.56	0.68
-n2	0.85	0.83	0.84	0.70	0.66	0.84	0.73	0.71	0.75	0.56	0.47	0.68
-n2-r5	0.87	0.61	0.84	0.72	0.59	0.84	0.76	0.58	0.75	0.56	0.59	0.68
-n2-r10	0.90	0.57	0.84	0.75	0.58	0.84	0.76	0.57	0.75	0.57	0.56	0.68
-n2-r15	0.95	0.50	0.84	0.76	0.58	0.84	0.76	0.58	0.75	0.58	0.55	0.68
-n2-r20	0.95	0.48	0.84	0.80	0.57	0.84	0.77	0.55	0.75	0.59	0.54	0.68
-n1	1.34	1.34	0.84	1.05	1.03	0.84	1.00	1.01	0.75	0.75	0.55	0.68
-n1-r5	1.31	0.06	0.84	1.02	0.02	0.84	0.96	0.00	0.75	0.74	0.53	0.68
-n1-r10	1.30	0.05	0.84	1.03	0.01	0.84	0.95	0.01	0.75	0.74	0.52	0.68
-n1-r15	1.30	0.14	0.84	1.03	0.01	0.84	0.95	0.02	0.75	0.74	0.52	0.68
-n1-r20	1.32	0.17	0.84	1.02	0.07	0.84	0.96	0.00	0.75	0.74	0.51	0.68
-k0.1	0.74	0.67	0.84	0.58	0.51	0.84	0.63	0.58	0.75	0.48	0.77	0.68
-k0.1-r5	0.77	0.69	0.84	0.70	0.62	0.84	0.63	0.68	0.75	0.51	0.00	0.68
-k0.1-r10	0.79	0.67	0.84	0.70	0.61	0.84	0.63	0.68	0.75	0.52	0.00	0.68
-k0.1-r15	0.80	0.67	0.84	0.72	0.59	0.84	0.67	0.66	0.75	0.53	0.00	0.68
-k0.1-r20	0.82	0.64	0.84	0.72	0.59	0.84	0.68	0.67	0.75	0.53	0.00	0.68
-k0.2	0.76	0.68	0.84	0.57	0.51	0.84	0.62	0.57	0.75	0.48	0.44	0.68
-k0.2-r5	0.78	0.67	0.84	0.70	0.61	0.84	0.62	0.69	0.75	0.51	0.59	0.68

-k0.2-r10	0.78	0.66	0.84	0.71	0.60	0.84	0.66	0.68	0.75	0.52	0.57	0.68
-k0.2-r15	0.80	0.64	0.84	0.72	0.59	0.84	0.67	0.67	0.75	0.53	0.56	0.68
-k0.2-r20	0.81	0.62	0.84	0.73	0.59	0.84	0.68	0.66	0.75	0.53	0.54	0.68
-k0.5	0.74	0.67	0.84	0.57	0.51	0.84	0.64	0.62	0.75	0.48	0.44	0.68
-k0.5-r5	0.76	0.70	0.84	0.66	0.62	0.84	0.64	0.65	0.75	0.51	0.59	0.68
-k0.5-r10	0.79	0.68	0.84	0.68	0.61	0.84	0.64	0.65	0.75	0.52	0.57	0.68
-k0.5-r15	0.81	0.65	0.84	0.69	0.59	0.84	0.69	0.65	0.75	0.53	0.56	0.68
-k0.5-r20	0.82	0.64	0.84	0.69	0.62	0.84	0.70	0.65	0.75	0.55	0.55	0.68
-k1.0	0.78	0.71	0.84	0.62	0.55	0.84	0.70	0.62	0.75	0.49	0.44	0.68
-k1.0-r5	0.83	0.62	0.84	0.71	0.58	0.84	0.69	0.60	0.75	0.51	0.58	0.68
-k1.0-r10	0.85	0.64	0.84	0.73	0.57	0.84	0.69	0.58	0.75	0.53	0.57	0.68
-k1.0-r15	0.90	0.57	0.84	0.75	0.56	0.84	0.71	0.57	0.75	0.54	0.56	0.68
-k1.0-r20	0.94	0.53	0.84	0.76	0.54	0.84	0.71	0.58	0.75	0.54	0.55	0.68