
Traversing chemical space with active deep learning for low-data drug discovery

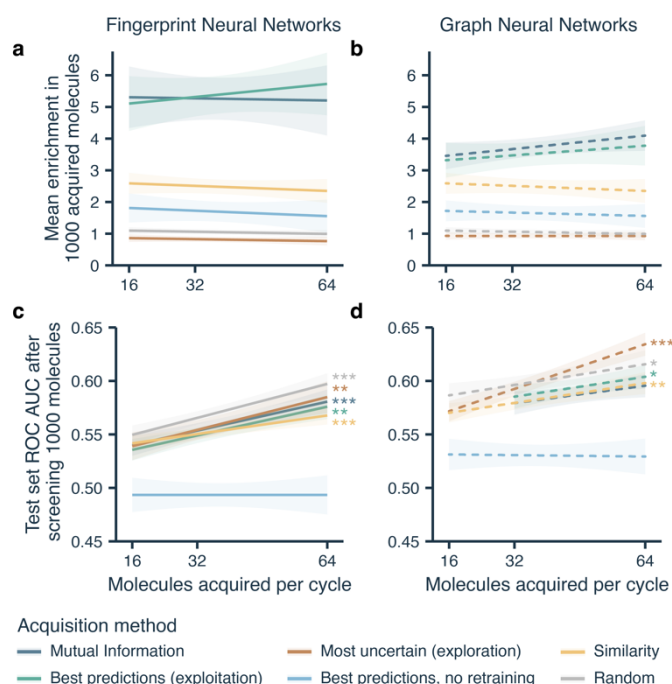
In the format provided by the
authors and unedited

Table of Contents

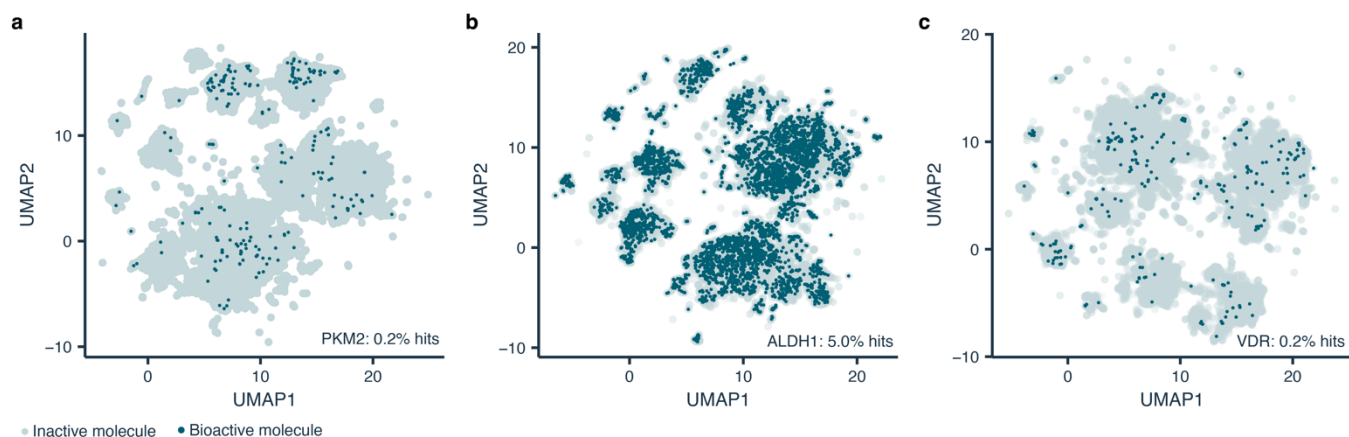
Preliminary results on the effects of the acquisition size	2
Supplementary Figure 1: The effects of the acquisition size on hit enrichment and predictive performance.	2
Supplementary Figure 2: Visualization of the chemical space of the selected virtual screening libraries.....	3
Supplementary Figure 3: Hit similarity distribution in the selected virtual screening libraries.	3
Supplementary Figure 4: The effects of structural diversity in the start data on hit enrichment and predictive performance.	4
Supplementary Figure 5: Hit discovery of active learning with perturbed features.	5
Supplementary Figure 6: performance of different active learning strategies.....	6
Supplementary Figure 7: Substructure discovery in acquired molecules.....	7
Supplementary Figure 8: Effect of acquisition functions on hit discovery for differently sized starting sets.	8
Supplementary Figure 9: Effect of acquisition functions on hit discovery on six low-yield datasets.	9
Supplementary Figure 10: The relationship between hit discovery and library 'yield'.....	9
Supplementary Table 1: Comparison of selected active learning studies using experimentally obtained data.	10
Supplementary Table 2: Statistical differences between hit enrichment with different acquisition methods.....	10
Supplementary Table 3: Predictive performance of models with different acquisition functions over active learning cycles.....	11
Supplementary Table 4: Molecular substructures used for structural analysis and molecular graph featurization.	12
Supplementary Table 5: Summary of low-yield datasets.....	13
Molecular Descriptors.	13
References	13

Preliminary results on the effects of the acquisition size

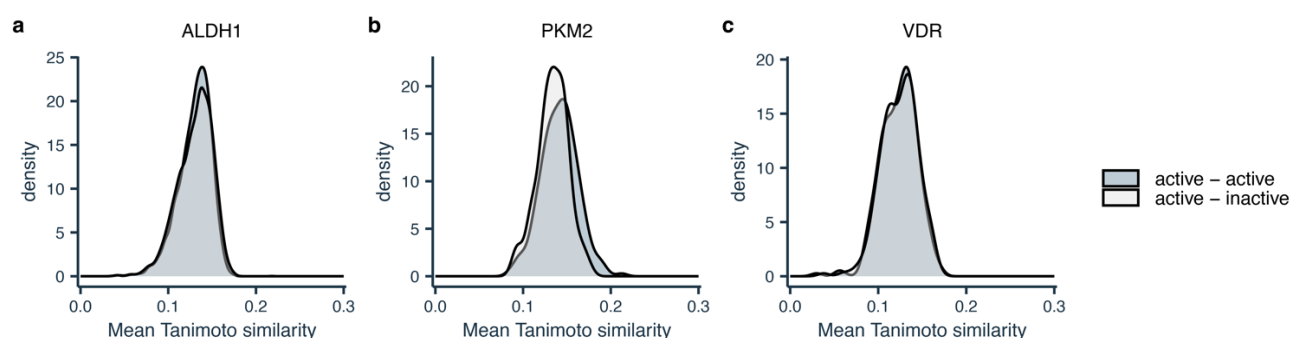
An important practical consideration is the number of molecules that are acquired and screened every cycle. A lower acquisition size would result in smaller screening steps. Acquiring fewer molecules per cycle results in more frequent model updates, potentially yielding better models, while simultaneously being lower throughput. To get some insights in the effect of the acquisition size, we compared acquiring either 16, 32, or 64 molecules for follow-up in each active learning cycle. We performed this experiment on only the ALDH1 dataset. We found that the enrichment factor remained similar regardless of acquisition size, showing no statistically significant differences (Supp. Fig. 1a, b). The same is true for model performance, as measured by the ROC AUC on the test set (Supp. Fig. 1c, d). In other data modalities, it is generally assumed that a lower acquisition size results in better performance, since models are more frequently updated and a smaller batch of samples has a lower likelihood of containing redundant information^{1,2}. However, when considering molecules, our results do not show this for the number of screening cycles performed in this study.



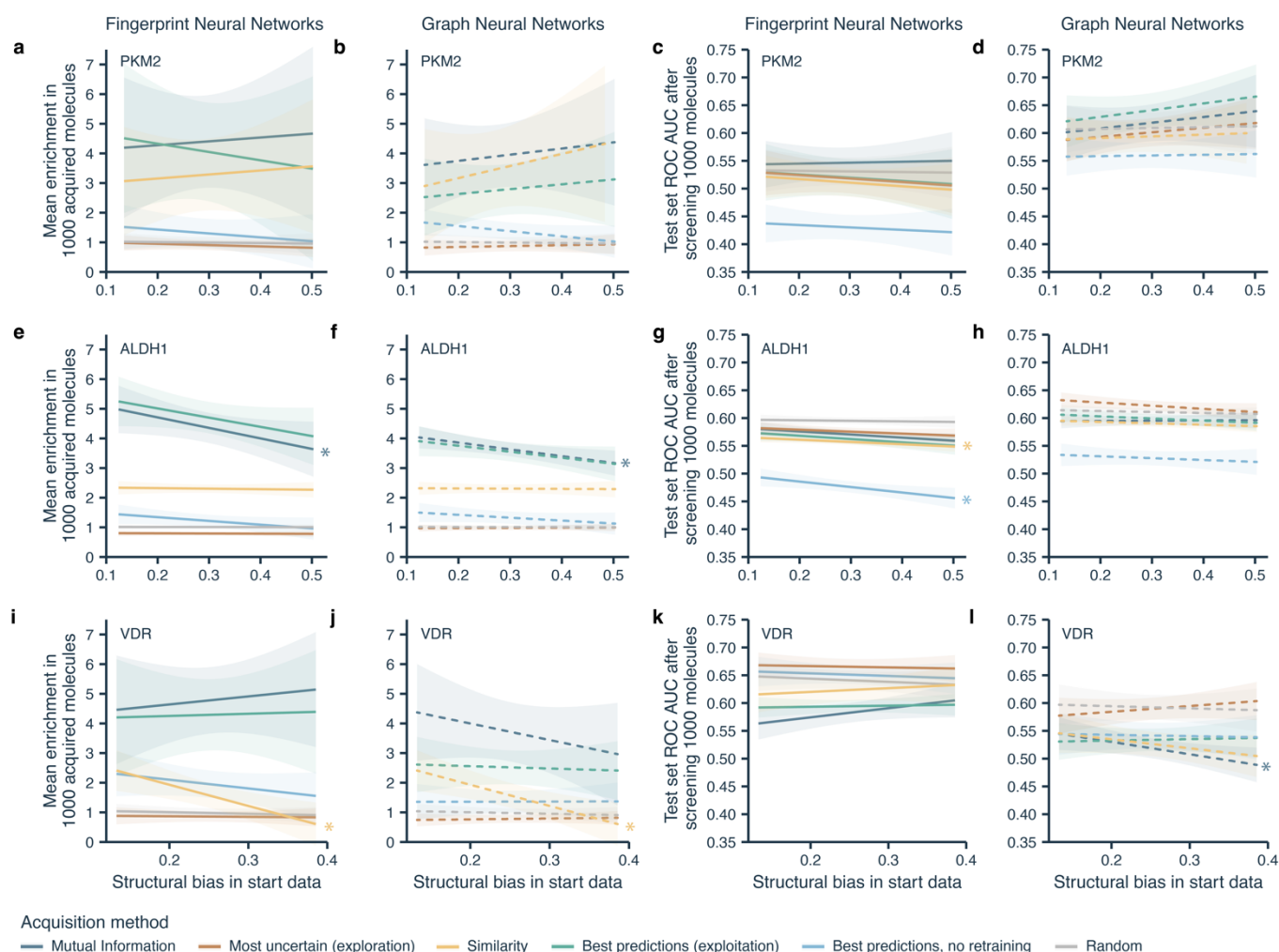
Supplementary Figure 1: The effects of the acquisition size on hit enrichment and predictive performance. Data is represented as the mean $\pm$ SEM of 10 experiments with different random starting molecules. Predictive performance is reported as the Area under the Receiver Operator Curve (ROC AUC) on the test set. Results are shown for models trained using ECFPs and molecular graphs on the ALDH1 dataset. Kendall's τ is used to test rank-based association, statistical significance is noted as * = $p < 0.05$, ** = $p < 0.001$, *** = $p < 0.0001$.



Supplementary Figure 2: Visualization of the chemical space of the selected virtual screening libraries. **a.** Pyruvate kinase M2 data (PKM2), containing 223 hits (agonists) out of 100,000 molecules (0.2%). **b.** Aldehyde dehydrogenase 1 (ALDH1), containing 4,986 hits (inhibitors) out of 100,000 molecules (5.0%). **c.** Vitamin D receptor data (VDR), containing 239 hits (agonists) out of 100,000 molecules (0.2%). Here, all molecules in the each dataset ($n=100,000$) are projected into 2D from a set of molecular descriptors using UMAP³ (see *Materials and Methods*). Every point represents a molecule in the dataset, with hit molecules highlighted in dark blue. The closer the molecules, the more similar their physicochemical properties.

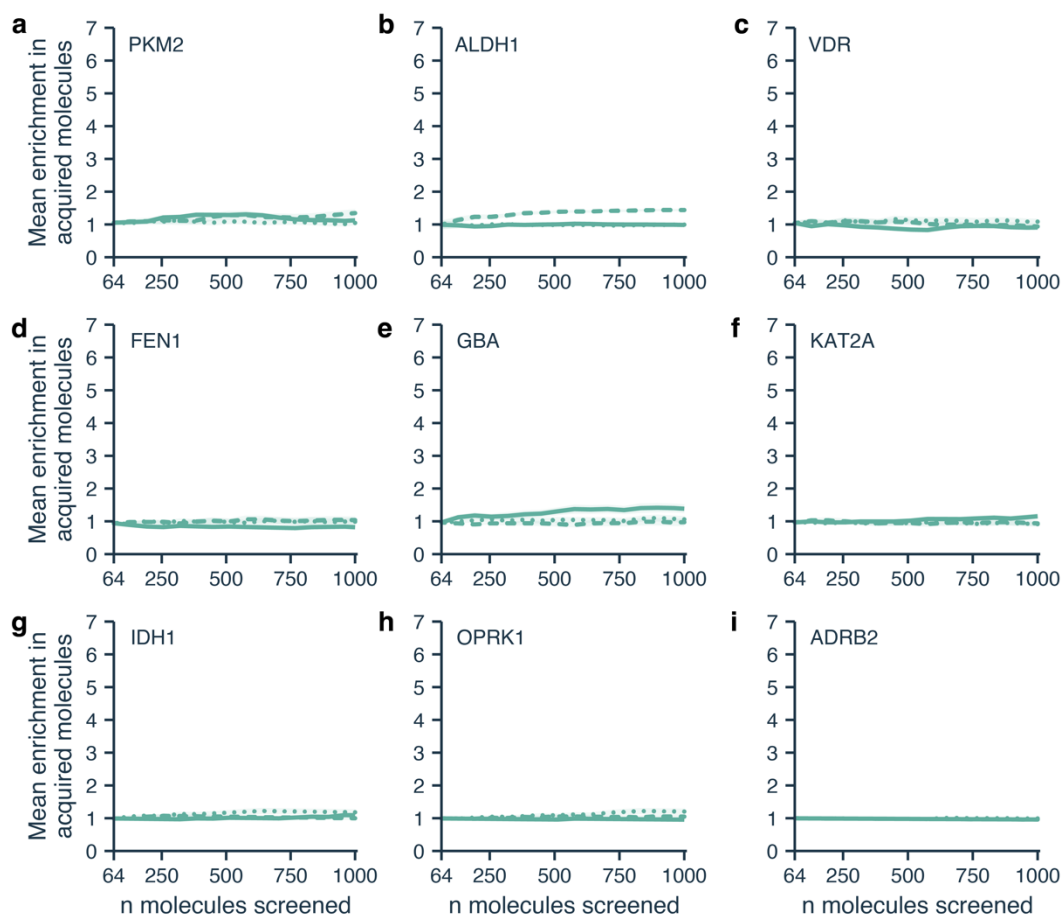


Supplementary Figure 3: Hit similarity distribution in the selected virtual screening libraries. **a.** Pyruvate kinase M2 data (PKM2), containing 223 hits (agonists) out of 100,000 molecules (0.2%). **b.** Aldehyde dehydrogenase 1 (ALDH1), containing 4,986 hits (inhibitors) out of 100,000 molecules (5.0%). **c.** Vitamin D receptor data (VDR), containing 239 hits (agonists) out of 100,000 molecules (0.2%). Here, the mean Tanimoto similarity between hits and between hits and inactives is computed. Similarity within actives is not remarkably different from the similarity between active and inactive molecules, indicating that hit chemotypes are not overrepresented.



Supplementary Figure 4: The effects of structural diversity in the start data on hit enrichment and predictive performance.

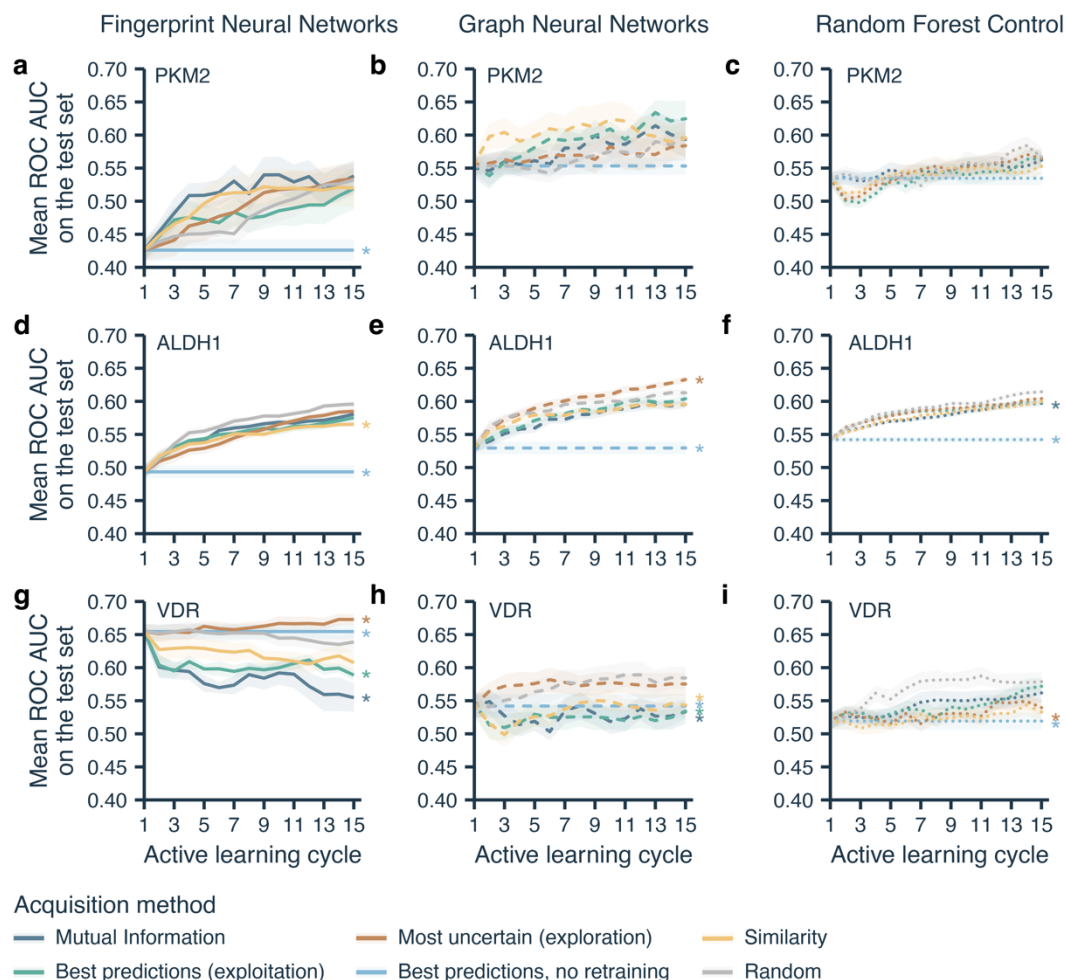
Predictive performance is reported as the Area under the Receiver Operator Curve (ROC AUC) on the test set. Diversity is reported as the mean Tanimoto similarity of Extended Connectivity Fingerprints⁴ between all molecules in the starting set. The higher the mean similarity, the lower the diversity. Kendalls τ is used to test rank-based association, statistical significance is noted as * = $p < 0.05$. Data is represented as the mean $\pm$ SEM of 10 experiments with different random starting molecules. Molecular fingerprints (ECFPs) are reported as solid lines, and graph neural networks as dashed lines. Every row represents one dataset: **a, b, c, d.** PKM2, **e, f, g, h.** ALDH1, **i, j, k, l.** VDR.



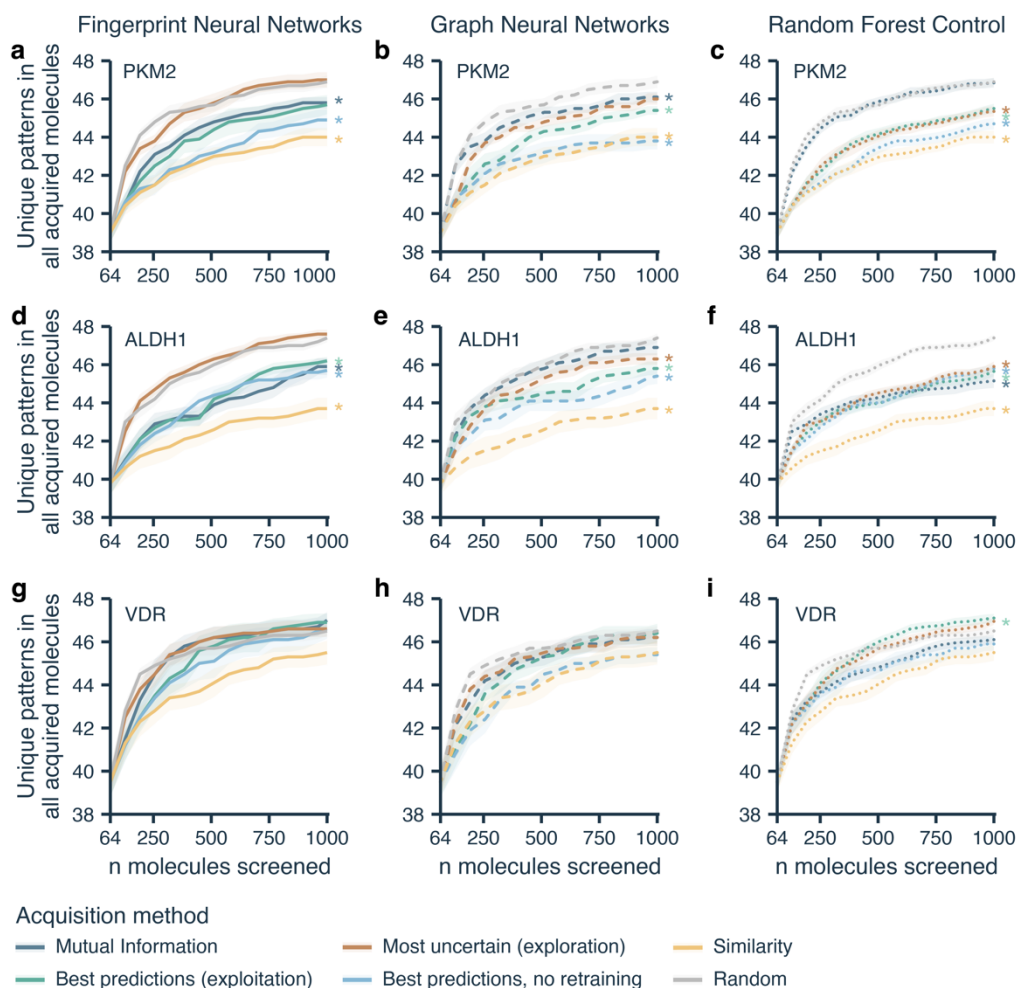
Method

— Fingerprint Neural Networks - - Graph Neural Networks . . . Random Forest Control

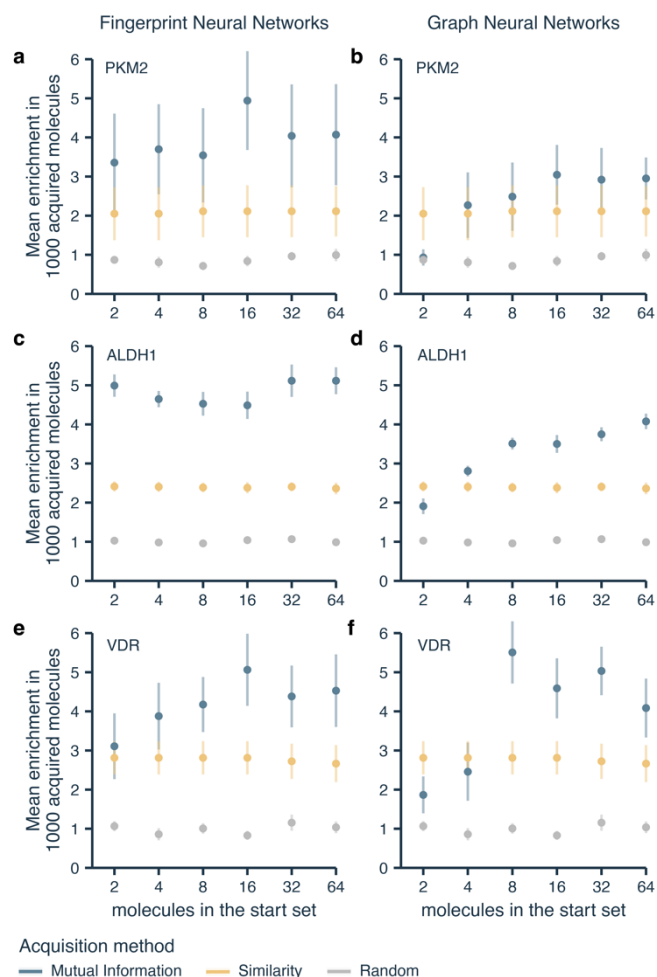
Supplementary Figure 5: Hit discovery of active learning with perturbed features. Enrichment factor of all acquired molecules with models trained using randomly perturbed molecular fingerprints (ECFPs) and molecular graphs is reported across nine different datasets. Molecules are acquired using best predictions (exploitation). Data is represented as the mean $\pm$ SEM of 10 experiments with different random starting molecules. All data is perturbed once at the start, so features stay consistent across active learning cycles. Additionally, for each experiment, three perturbations with different random seeds are performed. Molecular fingerprint neural networks are reported as solid lines, graph neural networks as dashed lines, and random forest control models using ECFPs as dotted lines. **a.** Pyruvate kinase M2 (PKM2), **b.** Aldehyde dehydrogenase 1 (ALDH1), **c.** Vitamin D receptor (VDR), **d.** FLAP endonuclease (FEN1), **e.** Glucocerebrosidase (GBA), **f.** Lysine acetyltransferase 2A (KAT2A), **g.** Isocitrate dehydrogenase (IDH1), **h.** Kappa opioid receptor (OPRK1), **i.** Beta2 adrenergic receptor (ADRB2).



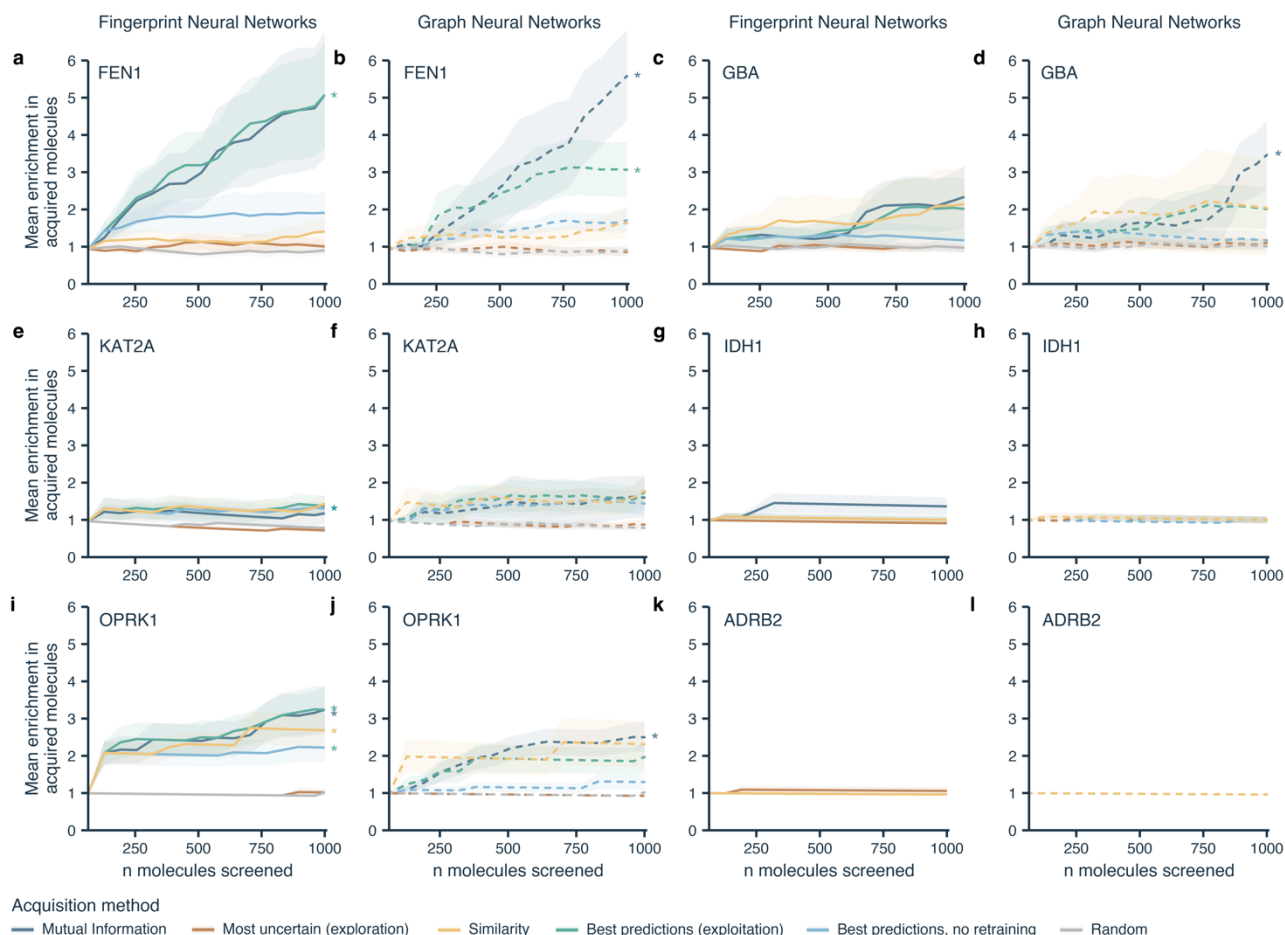
Supplementary Figure 6: performance of different active learning strategies. Predictive performance is reported as the Area under the Receiver Operator Curve (ROC AUC) on the test set. The mean ROC AUC over cycles is shown from machine/deep learning models trained using ECFPs and molecular graphs across three different datasets. Data is represented as the mean $\pm$ SEM of 10 experiments with different random starting molecules. Molecular fingerprints (ECFPs) are reported on the left-hand column (solid line), graph neural networks on the middle column (dashed line), and a random forest baseline model using ECFPs on the right-hand columns (dotted line). Every row represents one dataset: **a, b.** PKM2, **c, d.** ALDH1, **e, f.** VDR. Two-sided, paired, Wilcoxon Rank Sum tests are performed without multiple testing correction on the last cycle between random acquisition (grey) and the other acquisition methods. Statistical significance ($p < 0.05$) is denoted as *.



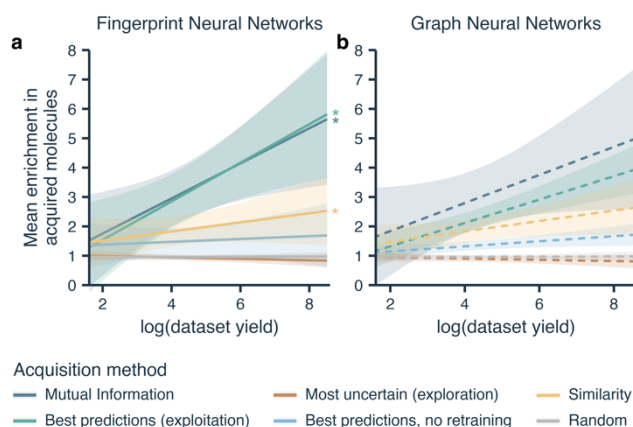
Supplementary Figure 7: Substructure discovery in acquired molecules. Molecular fingerprints (ECFPs) in combination with neural networks are reported on the left-hand column (solid line), graph neural networks on the middle column (dashed line), and ECFPs with random forest in the right-hand column (dotted line). Every row represents one dataset: **a, b, c.** PKM2, **d, e, f.** ALDH1, **g, h, i.** VDR. Data is represented as the mean $\pm$ SEM of 10 experiments with different random starting molecules, 64 molecules are acquired each cycle. Two-sided, paired, Wilcoxon Rank Sum tests are performed without multiple testing correction on the last cycle between random acquisition (grey) and the other acquisition methods. Statistical significance ($p < 0.05$) is denoted as *.



Supplementary Figure 8: Effect of acquisition functions on hit discovery for differently sized starting sets. Enrichment factor of all acquired molecules with models trained using ECFPs and molecular graphs, reported across three different datasets. Data is represented as the mean $\pm$ SEM of 10 experiments with different random starting molecules. Molecular fingerprints (ECFPs) are reported on the left-hand column, and graph neural networks on the right-hand column. Every row represents one dataset: **a, b**. PKM2, **c, d**. ALDH1, **e, f**. VDR.



Supplementary Figure 9: Effect of acquisition functions on hit discovery on six low-yield datasets. Enrichment factor of all acquired molecules with deep learning models trained using ECFPs and molecular graphs, reported across six different datasets from LIT-PCBA⁵ containing few hit molecules. Data is represented as the mean $\pm$ SEM of 10 experiments with different random starting molecules. Molecular fingerprints (ECFPs) are reported on the first and third column (solid line), and graph neural networks on the second and fourth column (dashed line). Datasets are shown by order of the number discoverable hits present in the full dataset ($n=100,000$, see Sup. Table 4): **a, b.** FLAP endonuclease (FEN1), 100 hits, **c, d.** Glucocerebrosidase (GBA), 55 hits, **e, f.** Lysine acetyltransferase 2A (KAT2A), 54 hits, **g, h.** Isocitrate dehydrogenase (IDH1), 11 hits, **i, j.** Kappa opioid receptor (OPRK1), 9 hits, **k, l.** Beta2 adrenergic receptor (ADRB2), 5 hits. Two-sided, paired, Wilcoxon Rank Sum tests are performed without multiple testing correction on the last cycle between random acquisition (grey) and the other acquisition methods. Statistical significance ($p < 0.05$) is denoted as *.



Supplementary Figure 10: The relationship between hit discovery and library 'yield'. Enrichment factor of all acquired molecules at the last active learning cycle with deep learning models trained using ECFPs and molecular graphs, reported across nine different datasets from LIT-PCBA⁵ containing different numbers of discoverable hit molecules ('yield'). Datasets included from high to low yield are ALDH1 (4,986 hits), VDR (239 hits), PKM2 (223 hits), FEN1 (100 hits), GBA (55 hits), KAT2A (54 hits), IDH1 (11 hits), OPRK1 (9 hits), ADRB2 (5 hits). Data is represented as the mean $\pm$ SEM of 10 experiments with different random starting molecules. Results are shown for neural networks trained using ECFPs and molecular graphs. Kendall's τ is used to test rank-based association, with * noting $p < 0.05$.

Supplementary Table 1: Comparison of selected active learning studies using experimentally obtained data. Studies using docking or simulation-based data are not mentioned. Reker and Schneider⁶ use a bioactivity dataset of 3,780 molecules, which they split into three equal parts (start/library/test). They report prediction error, distribution of bioactivity, and number of molecular scaffolds discovered over cycles. Zhang and Lee⁷ use six datasets: Melting point (melting temperature), 3,025 molecules; ESOL (solubility), 1,128 molecules; FreeSolv (hydration free energy), 643 molecules; CatS (bioactivity), 595 molecules; Malaria (bioactivity), 13,417 molecules; p450 (bioactivity), 8,817 molecules, of which they use 20% as test sets, 25% as starting set and the remaining as screening library. They report the prediction error over cycles. Bellamy *et al.*⁸ uses 288 ChEMBL datasets (bioactivity) ranging from 801 to 5,742 molecules and two PubChem bioactivity datasets with 8,817 and 13,417 molecules. They use 100 molecules as the starting set and the remainder as the screening library and report enrichment factor and hits discovered. Li and Rangerajan⁹ use four datasets: QM7 (enthalpy of atomization), 7,102 molecules; PHYSPROP boiling point (boiling temperature), 5,434 molecules; PHYSPROP melting point (melting temperature), 8,698 molecules; PHYSPROP logP (logP), 13,402 molecules of which they use 400 molecules as starting point and the remainder as screening library. They track prediction error and highest acquired values. Gorantla *et al.*¹⁰ use four datasets: TYK2 (binding affinity), 9,997 molecules; USP7 (bioactivity), 4,535 molecules; D2R (bioactivity), 2,502 molecules; and Mpro (bioactivity), 665 molecules. They use 60 molecules as the starting set and the remainder as the screening library and report recall and F1 score. Pyzer-Knapp¹¹ uses two datasets: MMP12 (bioactivity), 1,880 molecules and Malaria (bioactivity), 18,924 molecules. They start with 20 or 3 molecules respectively and use the remainder as screening library. They report the maximum discovered values, the mean of the top 10%, and hits discovered. Fujiwara *et al.*¹² use an annotated screening bioactivity library of 240,000 molecules, where they focus on three targets. They start with either 200 or 2,000 molecules, include similarity-based acquisition and report hits discovered. Reker *et al.*¹³ use one dataset: CXCR4 (bioactivity) containing 287 molecules, to perform prospective screening on 1,465,960 molecules. They report ligand efficiency and bioactivity. Rakers *et al.*¹⁴ use three datasets: KS5 (bioactivity), 39,706 molecules; GS3 (bioactivity), 47,602 molecules; GLASS (bioactivity), 69,960 molecules. They start with 2 random molecules and use the remainder as screening library. They report MCC and hits discovered.

Study	No. Datasets	Type of data	Train/starting set size	Library size	Approximate library/train ratio	No. acquisition functions
Reker and Schneider ⁶	1	Bioactivity	1,260	1,260	1	3
Zhang and Lee ⁷	6	Bioactivity/physicochemical	149 – 3,354	327 – 7,379	2	1
Bellamy <i>et al.</i> ⁸	290	Bioactivity	100	801 – 5,742	8 - 57	5
Li and Rangerajan ⁹	4	Physicochemical	400	5,434 – 13,402	14 - 34	2
Gorantla <i>et al.</i> ¹⁰	4	Bioactivity/binding affinity	60	655 – 9,997	11 – 167	4
Pyzer-Knapp ¹¹	2	Bioactivity	3/20	1,880/18,924	626/946	4
Fujiwara <i>et al.</i> ¹²	3	Bioactivity	200/2,000	240,000	1200/120	4
Reker <i>et al.</i> ¹³ (prospective)	1	Bioactivity	287	1,465,960	5,108	2-in-1
Rakers <i>et al.</i> ¹⁴	3	Bioactivity	2	39,706 – 69,960	34,980	2
Us	3	Bioactivity	2 - 64	100,000	50,000 – 1,563	5

Supplementary Table 2: Statistical differences between hit enrichment with different acquisition methods. Two-sided, paired, Wilcoxon Rank Sum tests are performed without multiple testing correction on the last cycle for the PKM2, ALDH1, and VDR datasets between different acquisition methods. Results are compared per acquisition method for hit enrichment after screening 1000 molecules across different machine/deep learning approaches: fingerprint neural networks (MLP), graph neural networks (GCN), and a random forest control using fingerprints (RF). Statistically significant differences ($p < 0.05$) are denoted as *.

	Model type	Best predictions (exploitation)			Best predictions, no retraining			Mutual Information			Most uncertain (exploration)		
		MLP	GCN	RF	MLP	GCN	RF	MLP	GCN	RF	MLP	GCN	RF
PKM2	MLP						*						*
	GCN				*					*			*
	RF								*		*	*	
ALDH1	MLP		*				*		*	*		*	*
	GCN	*		*			*	*		*	*		*
	RF		*		*	*		*	*		*	*	
VDR	MLP			*						*			*
	GCN			*						*			*
	RF	*	*					*	*		*	*	

Supplementary Table 3: Predictive performance of models with different acquisition functions over active learning cycles.

Performance is reported as Receiver Operator Curve Area Under the Curve (ROC AUC) on an independent test set (n=20,000). An acquisition size of 64 was used.

Acquisition function		PKM2			ALDH1			VDR		
		Cycle 1	Cycle 8	Cycle 15	Cycle 1	Cycle 8	Cycle 15	Cycle 1	Cycle 8	Cycle 15
MLPs + ECFPs	Mutual Information	0.426 ± 0.05	0.511 ± 0.08	0.538 ± 0.07	0.493 ± 0.03	0.562 ± 0.02	0.581 ± 0.01	0.655 ± 0.04	0.589 ± 0.06	0.555 ± 0.07
	Exploit	0.426 ± 0.05	0.474 ± 0.07	0.519 ± 0.10	0.493 ± 0.03	0.556 ± 0.02	0.575 ± 0.02	0.655 ± 0.04	0.599 ± 0.03	0.589 ± 0.03
	Exploit (no retrain)	0.426 ± 0.05	0.426 ± 0.05	0.426 ± 0.05	0.493 ± 0.03	0.493 ± 0.02	0.493 ± 0.03	0.655 ± 0.04	0.655 ± 0.04	0.655 ± 0.04
	Explore	0.426 ± 0.05	0.498 ± 0.08	0.534 ± 0.08	0.493 ± 0.03	0.552 ± 0.02	0.585 ± 0.02	0.655 ± 0.04	0.660 ± 0.03	0.673 ± 0.03
	Random	0.426 ± 0.05	0.475 ± 0.07	0.532 ± 0.07	0.493 ± 0.03	0.573 ± 0.01	0.596 ± 0.02	0.655 ± 0.04	0.652 ± 0.03	0.639 ± 0.04
	Similarity	0.426 ± 0.05	0.513 ± 0.09	0.520 ± 0.09	0.493 ± 0.03	0.551 ± 0.02	0.565 ± 0.01	0.655 ± 0.04	0.626 ± 0.04	0.608 ± 0.05
GNNs	Mutual Information	0.550 ± 0.05	0.580 ± 0.07	0.593 ± 0.09	0.531 ± 0.03	0.580 ± 0.02	0.596 ± 0.02	0.541 ± 0.06	0.542 ± 0.07	0.533 ± 0.06
	Exploit	0.554 ± 0.04	0.594 ± 0.04	0.625 ± 0.09	0.531 ± 0.03	0.588 ± 0.01	0.604 ± 0.02	0.546 ± 0.06	0.526 ± 0.05	0.534 ± 0.04
	Exploit (no retrain)	0.553 ± 0.04	0.553 ± 0.04	0.553 ± 0.04	0.530 ± 0.03	0.530 ± 0.03	0.530 ± 0.03	0.542 ± 0.06	0.542 ± 0.06	0.542 ± 0.06
	Explore	0.556 ± 0.05	0.570 ± 0.06	0.584 ± 0.07	0.531 ± 0.03	0.606 ± 0.02	0.633 ± 0.02	0.543 ± 0.06	0.572 ± 0.05	0.575 ± 0.05
	Random	0.553 ± 0.04	0.564 ± 0.07	0.595 ± 0.08	0.530 ± 0.03	0.597 ± 0.01	0.613 ± 0.02	0.545 ± 0.06	0.576 ± 0.06	0.584 ± 0.05
	Similarity	0.556 ± 0.04	0.617 ± 0.08	0.596 ± 0.08	0.530 ± 0.03	0.586 ± 0.01	0.596 ± 0.02	0.544 ± 0.06	0.547 ± 0.05	0.543 ± 0.05
RF + ECFPs	Mutual Information	0.535 ± 0.05	0.552 ± 0.04	0.567 ± 0.04	0.542 ± 0.02	0.573 ± 0.01	0.584 ± 0.01	0.519 ± 0.06	0.546 ± 0.08	0.549 ± 0.07
	Exploit	0.535 ± 0.05	0.548 ± 0.06	0.565 ± 0.07	0.542 ± 0.02	0.586 ± 0.01	0.601 ± 0.01	0.519 ± 0.06	0.528 ± 0.06	0.569 ± 0.04
	Exploit (no retrain)	0.535 ± 0.05	0.535 ± 0.05	0.535 ± 0.05	0.542 ± 0.02	0.542 ± 0.02	0.542 ± 0.02	0.519 ± 0.06	0.519 ± 0.06	0.519 ± 0.06
	Explore	0.535 ± 0.05	0.563 ± 0.08	0.571 ± 0.06	0.542 ± 0.02	0.586 ± 0.01	0.603 ± 0.01	0.519 ± 0.06	0.525 ± 0.05	0.523 ± 0.04
	Random	0.535 ± 0.05	0.548 ± 0.07	0.567 ± 0.07	0.542 ± 0.02	0.591 ± 0.01	0.614 ± 0.01	0.519 ± 0.06	0.582 ± 0.07	0.579 ± 0.06
	Similarity	0.535 ± 0.05	0.540 ± 0.06	0.553 ± 0.05	0.542 ± 0.02	0.581 ± 0.02	0.600 ± 0.01	0.519 ± 0.06	0.521 ± 0.04	0.533 ± 0.07

Supplementary Table 4: Molecular substructures used for structural analysis and molecular graph featurization.
Substructures were identified through different SMILES Arbitrary Target Specification (SMARTS) patterns.

Substructure	SMARTS	Pattern recognition	Graph featurization
Aliphatic chain of at least 8 members	[AR0]~[AR0]~[AR0]~[AR0]~[AR0]~[AR0]~[AR0]~[AR0]	X	X
Bi-cyclicity	[\$([*R2]([*R])([*R])([*R]))].[\$([*R2]([*R])([*R])([*R]))]	X	X
Ortho-substituted ring	*-!::aa-!:	X	X
Meta-substituted ring	*-!::aaa-!:	X	X
Para-substituted ring	*-!::aaaa-!:	X	X
Macrocyclic group	[r;!r3;!r4;!r5;!r6;!r7]	X	X
Spiro-ring centre	[X4;R2;r4,r5,r6]([@r4,r5,r6])([@r4,r5,r6])([@r4,r5,r6])	X	X
Cis or trans double bond in a ring	*/,[R]=,@[R]/,*	X	X
Unfused benzene ring	[cR1]1[cR1][cR1][cR1][cR1][cR1]1	X	X
Multiple non-fused benzene rings	[cR1]1[cR1][cR1][cR1][cR1][cR1]1.[cR1]1[cR1][cR1][cR1][cR1][cR1]1	X	X
Fused benzene rings	c12cccc1cccc2	X	X
Carbonyl group	[\$([CX3]=[OX1]),\$([CX3+]-[OX1-])]	X	X
Aldehyde	[CX3H1](=O)[#6]	X	X
Amide	[NX3][CX3](=[OX1])[#6]	X	X
Carbamate	[NX3,NX4+][CX3](=[OX1])[OX2,OX1-]	X	X
Carboxylate ion	[CX3](=O)[O-]	X	X
Carbonic acid or carbonic ester	[CX3](=[OX1])(O)O	X	X
Carboxylic acid	[CX3](=O)[OX1H0-,OX2H1]	X	X
Ester (also hits anhydrides)	[#6][CX3](=O)[OX2H0][#6]	X	X
Ketone	[#6][CX3](=O)[#6]	X	X
Ether	[OD2]([#6])[#6]	X	X
Mono-hydrogenated cation	[+H]	X	X
Amidinium	[NX3][CX3]=[NX3+]	X	X
Cyanamide	[NX3][CX2][NX1]	X	X
Primary or secondary amine	[NX3;H2,H1;!\$(NC=O)]	X	X
Enamine	[NX3]\$(C=C),\$(cc)]	X	X
Azole	[\$([nr5]:[nr5,or5,sr5]),\$([nr5]:[cr5]:[nr5,or5,sr5])]	X	X
Hydrazine	[NX3][NX3]	X	X
Hydrazone	[NX3][NX2]=[*]	X	X
Imine (substituted or un-substituted)	[\$([CX3]([#6])[#6]),\$([CX3H][#6])]=\$([NX2][#6]),\$([NX2H])]	X	X
Iminium	[NX3+]=[CX3]	X	X
Unsubstituted dicarboximide	[CX3](=[OX1])[NX3H][CX3](=[OX1])	X	X
Substituted dicarboximide	[CX3](=[OX1])[NX3H0]([#6])[CX3](=[OX1])	X	X
Nitrate group	[\$([NX3](=[OX1])(=[OX1])O),\$([NX3+](=[OX1-])(=[OX1])O)]	X	X
Nitrile	[NX1][CX2]	X	X
Nitro group	[\$([NX3](=O)=O),\$([NX3+](=O)[O-])][!#8]	X	X
Hydroxyl	[OX2H]	X	X
Enol	[OX2H][#6X3]=[#6]	X	X
Phenol	[OX2H][cX3]:[c]	X	X
Carbo-thioester	S([#6])[CX3](=O)[#6]	X	X
Thiol	[SX2]	X	X
Thioamide	[NX3][CX3]=[SX1]	X	X
Sulphide	[#16X2H0]	X	X
Mono-sulphide	[#16X2H0][!#16]	X	X
Two sulphides	[#16X2H0][!#16].[#16X2H0][!#16]	X	X
Sulfone	[\$([#16X4](=[OX1])=[OX1]),\$([#16X4+2]([OX1-])(OX1-)))]	X	X
Sulphonamide	[\$([SX4](=[OX1])(=[OX1])([O])[NX3]),\$([SX4+2]([OX1-])(OX1-)))([O])[NX3]]	X	X
Sulfoxide	[\$([#16X3]=[OX1]),\$([#16X3+][OX1-])]	X	X
Halogen	[F,Cl,Br,I]	X	X
Three halide group	[F,Cl,Br,I].[F,Cl,Br,I].[F,Cl,Br,I]	X	X
Rotatable bond	[!\$(*)&!D1]-!@!\$(*)&!D1]		X
Acylic-bonds	*!@*		X
Specified chiral carbon	[\$([#6X4@]([*])([*])([*]),\$([#6X4@H]([*])([*])([*]))]		X
Quaternary nitrogen	[\$([NX4+]),\$([NX4]=*)]		X
Non-ring atom	[!R]		X
Ring atom	[R]		X
Any carbon attached to any halogen	[r;!r3;!r4;!r5;!r6;!r7]		X
Triply bonded nitrogen	[\$([NX1]#*)]		X
Cis/trans double/aromatic ring bond	*/,[R]=,::@[R]/,*		X
Hydroxyl in alcohol	[#6][OX2H]		X
Thio analogue of carbonyl	[#6X3](=[SX1])([!N])[!N]		X
Divalent oxygen	[\$([OX2])]		X
Isolating carbon	[\$([#6+0]),!\$(C(F)(F)F),!\$(c(:[!c]):[!c]):!\$([#6]=,[!#6])]		X
Single bond and not in a ring	*-!@*		X
S in aromatic 5-ring with lone pair	[sX2r5]		X
Aromatic 5-ring O with lone pair	[oX2r5]		X
N in 5-membered aromatic ring	[\$([nX2r5]:[a-]),\$([nX2r5]:[a]:[a-])]		X

Supplementary Table 5: Summary of low-yield datasets. Subsets of six datasets from LIT-PCBA⁵ were used: FLAP endonuclease (FEN1), Glucocerebrosidase (GBA), Lysine acetyltransferase 2A (KAT2A), Isocitrate dehydrogenase (IDH1), Kappa opioid receptor (OPRK1), and Beta2 adrenergic receptor (ADRB2). Mean and standard deviations of the number of hit molecules in the 10 starting sets are reported. An external test set of 20,000 molecules served to monitor performance over cycles.

Dataset	Screening library size	Test set size	Hits in starting sets
FEN1	100,000 (100 hits, 0.10%)	20,000 (20 hits, 0.10%)	1.00 ± 0.00
GBA	100,000 (55 hits, 0.06%)	20,000 (11 hits, 0.06%)	1.00 ± 0.00
KAT2A	100,000 (54 hits, 0.05%)	20,000 (11 hits, 0.06%)	1.00 ± 0.00
IDH1	100,000 (11 hits, 0.01%)	20,000 (2 hits, 0.01%)	1.00 ± 0.00
OPRK1	100,000 (9 hits, 0.01%)	20,000 (2 hits, 0.01%)	1.00 ± 0.00
ADRB2	100,000 (5 hits, 0.01%)	20,000 (1 hits, 0.01%)	1.00 ± 0.00

Molecular Descriptors. For the purpose of chemical space visualization (Fig. 4 and Supp. Fig. 2), all available molecular descriptors in RDKit⁴ (v. 2022.09.5) were computed for each molecule: *MaxAbsEStateIndex*, *MaxEStateIndex*, *MinAbsEStateIndex*, *MinEStateIndex*, *qed*, *SPS*, *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*, *NOCCount*, *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_Iimine*, *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_Topliiss*, *fr_lactam*, *fr_lactone*, *fr_methoxy*, *fr_morpholine*, *fr_nitrile*, *fr_nitro*, *fr_nitro_omom*, *fr_nitro_omom_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*.

References

- Kirsch, A., van Amersfoort, J. & Gal, Y. BatchBALD: Efficient and Diverse Batch Acquisition for Deep Bayesian Active Learning. in *Advances in Neural Information Processing Systems* (eds. Wallach, H. et al.) vol. 32 (Curran Associates, Inc., 2019).
- Citovsky, G. *et al.* Batch Active Learning at Scale. in *Advances in Neural Information Processing Systems* (eds. Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P. S. & Vaughan, J. W.) vol. 34 11933–11944 (Curran Associates, Inc., 2021).
- McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. Preprint at <https://doi.org/10.48550/arXiv.1802.03426> (2020).
- Rogers, D. & Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **50**, 742–754 (2010).

5. Tran-Nguyen, V.-K., Jacquemard, C. & Rognan, D. LIT-PCBA: An Unbiased Data Set for Machine Learning and Virtual Screening. *J. Chem. Inf. Model.* **60**, 4263–4273 (2020).
6. Reker, D. & Schneider, G. Active-learning strategies in computer-assisted drug discovery. *Drug Discovery Today* **20**, 458–465 (2015).
7. Zhang, Y. & Lee, A. A. Bayesian semi-supervised learning for uncertainty-calibrated prediction of molecular properties and active learning. *Chem. Sci.* **10**, 8154–8163 (2019).
8. Bellamy, H., Rehim, A. A., Orhobor, O. I. & King, R. Batched Bayesian Optimization for Drug Design in Noisy Environments. *J. Chem. Inf. Model.* **62**, 3970–3981 (2022).
9. Li, B. & Rangarajan, S. A diversity maximizing active learning strategy for graph neural network models of chemical properties. *Mol. Syst. Des. Eng.* **7**, 1697–1706 (2022).
10. Gorantla, R., Kubincová, A., Suutari, B., Cossins, B. P. & Mey, A. S. J. S. Benchmarking Active Learning Protocols for Ligand-Binding Affinity Prediction. *J. Chem. Inf. Model.* **64**, 1955–1965 (2024).
11. Pyzer-Knapp, E. O. Bayesian optimization for accelerated drug discovery. *IBM J. Res. & Dev.* **62**, 2:1-2:7 (2018).
12. Fujiwara, Y. *et al.* Virtual Screening System for Finding Structurally Diverse Hits by Active Learning. *J. Chem. Inf. Model.* **48**, 930–940 (2008).
13. Reker, D., Schneider, P. & Schneider, G. Multi-objective active machine learning rapidly improves structure–activity models and reveals new protein–protein interaction inhibitors. *Chem. Sci.* **7**, 3919–3927 (2016).
14. Rakers, C., Reker, D. & Brown, J. B. Small Random Forest Models for Effective Chemogenomic Active Learning. *J. Comput. Aided Chem.* **18**, 124–142 (2017).