Multi-fidelity machine learning models for improved high-throughput screening predictions

David Buterez[1] Jon Paul Janet[2] Steven J. Kiddle [3] Pietro Liò[1]
db804 @ cam.ac.uk

Abstract

High throughput screening (HTS) is one of the leading techniques for hit identification in drug discovery and is often done in two phases, primary and confirmatory. The resulting data is multi-fidelity, with noisy primary screening data available on a large number of compounds and higher quality confirmatory data on a low-to-moderate number of compounds. Existing computational pipelines do not integrate primary and confirmatory screening data of individual HTS campaigns, resulting in millions of potentially useful screening data points being unused in models of confirmatory bioactivity prediction. Furthermore, there is currently a lack of publicly available multi-fidelity bioactivity benchmarks to support modelling real-world high-throughput screening data.

To address these challenges, we first compiled a public collection of 23 multi-fidelity HTS datasets from PubChem for benchmarking, including more than 6.1 million data points. Additionally, we assembled a private collection of 19 AstraZeneca HTS datasets, spanning more than 22.8 million data points. We then designed and evaluated machine learning models to assess the improvements offered by the integration of multi-fidelity data, including classical machine learning and novel deep learning approaches, the latter based on graph neural networks. Jointly modelling primary and confirmatory data led to a decrease of 12% in mean absolute error (MAE) and an increase of 152% in Pearson R^2 on the public datasets, and a reduction of 17% in MAE coupled with an uplift of 46% in R^2 on the AstraZeneca datasets (averaged across all evaluated methods). Furthermore, supplementing with molecular embeddings produced by previously trained deep learning models led to improved metrics for compounds that were not part of the primary screen, with up to double the baseline performance. We conclude that joint modelling of multi-fidelity HTS data improves predictive performance and that deep learning enables the use of unique and highly desirable strategies such as leveraging signals from multi-million scale datasets and transfer learning.

1 Introduction

High-throughput screening (HTS) consists of a set of largely-automated techniques to experimentally determine relevant biochemical interactions for large collections of synthetic compounds. The origins of HTS can be traced back to around three decades ago, when pharmaceutical companies started transitioning from natural products screening of up to 10,000 compounds per week to synthetic compound screening. The shift was enabled by progress in automated and parallel processing of microtitre plates, as well as rapidly-expanding compound library sizes resulting from combinatorial chemistry. Its popularity was also due to heightened interest in target-based drug discovery thanks to advances in molecular biology and genomics [PW07; Mof+17]. Initially, the brute-force approach of HTS, the presumed lack of quality, and the early lack of successful commercial drugs were criticised. However, HTS technology has matured enough to be widely accepted in industrial and academic settings, with a considerable number of FDA-approved drugs originating from high-throughput screens. [Mac+11]. Perola reported that 19 out of 58 drugs approved between 1991 and 2008 were derived from HTS campaigns [Per10], while a more recent analysis examined 66 clinical candidates reported in the *Journal of Medicinal Chemistry* between 2016 and 2017, determining that 29% of compounds were based on hits generated by large random compound library screening [BB18]. Recent estimates of modern compound libraries for some of the largest pharmaceutical companies indicate sizes varying from 1.2 million to 4 million molecules, and a throughput of more than 100,000 screened compounds per day for leading laboratories, on a variety of assays [Vol+19]. Furthermore, there is evidence that a large and diverse chemical library, often achieved through sharing of proprietary libraries, in combination with multiple, parallel screening approaches and

[1] Department of Computer Science and Technology – University of Cambridge, UK
[2] Molecular AI, Discovery Sciences, Biopharmaceuticals R&D, AstraZeneca, Gothenburg, Sweden
[3] AI & Analytics, Data Science & AI, R&D, AstraZeneca, Cambridge, UK

cutting-edge laboratory techniques increased the success of AstraZeneca early-stage discovery projects, measured in terms of successful transitions to lead optimisation [Mor+18].

With high-throughput screening now producing millions of drug-protein interactions per project, the idea of a digital pipeline for data analysis and targeted compound generation has gained traction, especially when based on machine learning and modern 'Big Data' approaches. Applications of machine learning for modelling structure-property relationships of therapeutic interest have been known for more than two decades, coinciding with the time when molecular datasets were first becoming available and machine learning algorithms were gaining popularity. Notable historical examples include the application of Support Vector Machines [BGV92] (SVM) on public [Bur+01] and GlaxoSmithKline [TH03] data, as well as Random Forest [Bre01] (RF) models [Sve+04] on public data, both algorithms representing the state-of-the-art at the time. It is worth noting that this category of algorithms (referred to as 'classical' or 'shallow' models in this paper) is still competitive and widely used today. For example, a recent review from Boehringer Ingelheim reported SVM and RF as the top performers on internal ADMET (absorption, distribution, metabolism, excretion, and toxicity) prediction tasks, and furthermore presented mixed results regarding the benefit of increasing training set sizes, with 8 out of 23 datasets recording negative effects on the predictive performance when using more data [ASB]. The debate of whether more data is helpful for virtual screening, usually framed in the context of virtual docking, has still not been settled [Cla20]. Bayer's *in silico* ADMET platform currently uses a mix of classical and deep learning strategies, emphasising that the classification performance is identical between established (SVM, RF) models and modern artificial neural network architectures, but due to certain strengths of deep learning such as scalability and multi-task learning, it is increasingly being used internally [Gö+20]. Nonetheless, classical algorithms are normally discussed and evaluated alongside the latest machine learning applications for molecular prediction, such as MoleculeNet [Wu+18] and ChemProp [Yan+19], and are even used independently in novel directions [KSL20].

During the last decade, several deep learning developments added another dimension to chemical modelling, accompanied by new expectations and hope. This accomplishment is mostly attributed to the success of convolutional neural networks and more recently to graph neural networks (GNN) [Duv+15], a framework that naturally accommodates the idea of a molecular graph, as well as other novel ideas such as SchNet [Sch+17], which used continuous-filter convolutional layers to model quantum interactions in molecules. In fields like computer vision, breakthroughs were only achieved after the development of high-quality, large-scale datasets, enabling the objective comparison of different architectures. Mirroring this approach, recent efforts led to an increased number of datasets and benchmarks dedicated to computational chemistry, such as MoleculeNet, LIT-PCBA [TNJR20] and Atom3D [Tow+20]. Furthermore, deep learning strategies such as few-shot learning [AT+17; Sta+21] and generative modelling [JBJ18] have started to be adapted to computational chemistry with success. Perhaps the most well-known recent example is the discovery of a broad-spectrum antibiotic that is structurally distant from conventional anti-bacterial compounds [Sto+20], named halicin, a breakthrough made possible by a directional message passing deep learning architecture.

Successfully exploiting the large amount of high-throughput screening data that is generated in the public and private domains is of great interest for the computational chemistry community. Notable contributions to the field include constructing historical HTS fingerprints [Pet+12; Hel+16; Lau+19; Stu+19] and attempting to directly solve the bioactivity classification problem with increasingly more sophisticated deep learning architectures [Yan+19; Gur+20]. However, certain limitations are not yet definitively addressed. In the case of HTS fingerprints, activity flags from hundreds of assays are assembled in a per-compound activity vector (fingerprint), with several studies showing that they outperform purely structural fingerprints in bioactivity prediction tasks. However, the strategy is inherently not scalable, since adding a single new compound would require screening it in hundreds of different assays. Furthermore, compounds might have problematically sparse representations and Laufkötter et al. report that a hybrid approach involving the molecular structure is preferable for both predictive performance and scaffold hopping capability [Lau+19]. On the other hand, tackling the bioactivity classification problem has the major obstacle of a massive class imbalance, since few molecules have truly favourable interactions with the protein target. Other recent advances focused on improving the brute-force approach of HTS by iterative virtual screening, where machine learning is used to design the next subset of compounds to be screened after screening an initial fraction of the library [Dre+21].

In this work, we propose and evaluate a methodology for exploiting large amounts of high-throughput screening data for bioactivity prediction, focusing on the multi-fidelity aspect of HTS. Most of the existing work reduces the task to a prediction problem with labels extracted from the highest-quality measurements available (*concentration* or *dose response*). However, this approach discards intermediary measurements, in particular millions of primary screening interactions (*single dose*), due to considerations of noise and uncertainty regarding appropriate integration methods. We hypothesise that leveraging the activity measured in the primary phase of HTS leads to more powerful quantitative structure-activity relationship (QSAR) models. This idea is motivated by the vastly larger chemical space (up to 3 orders of magnitude) covered in primary screens, and recent deep learning advances that are capable of learning relationships beyond simple similarity of fragments or entire molecules.

To help our investigation and motivate further research into this area, we first introduce a new collection of 23 multi-fidelity HTS datasets, assembled and filtered from PubChem assay data, out of which we examined a selection of 23 datasets, totalling more than 6.1 million unique interactions. This study is further supported by a set of 19 in-house AstraZeneca multi-fidelity datasets that we assembled, totalling over 22.8 million unique interactions. We validate our hypothesis that the integration of different data modalities is helpful by designing and evaluating a range of machine learning algorithms, including random forests (RF), support vector machines (SVM), and a novel specifically designed variational autoencoder architecture with graph convolutional layers (Figure 1). Furthermore, we employed our deep learning architecture to test the transfer learning potential, where previously-learnt molecular embeddings trained exclusively on primary screening data are leveraged by models of confirmatory data to improve predictions. Importantly, transfer learning enables predictions for compounds lacking primary screening measurements, as trained models can generate embeddings for previously unseen molecules. Our study also aims to clarify some long-standing doubts by quantifying the benefits of deep learning for improved high-throughput screening predictions and relating the size of the training datasets to model performance.

2 Design and Implementation

2.1 Multi-fidelity datasets

In this work, we define a *multi-fidelity HTS dataset* as a molecular dataset with two different experimentally-derived bioactivity measurements: single dose (SD) and dose response (DR). When referring to a multi-fidelity dataset, the identifier of the DR dataset is written first, followed by the SD identifier and separated by '–', e.g. AID2382 – AID2098. We use the term *primary* and *single dose* interchangeably; similarly for *confirmatory* and *dose response*.

The SD values are extracted from the primary screen, which evaluates a large library of compounds for activity at a single concentration. Most compounds which are recognised as active in the primary screen are further examined in the confirmatory screen, where the activities at multiple different concentrations, for each compound, are summarised in a single 'pXC50' activity value, with $\mathrm{X} \in \{\mathrm{I} = \text{inhibitory}, \mathrm{E} = \text{effective}, \mathrm{A} = \text{activity}\}$. The compounds with both SD and DR activities present represent a fraction of the entire compound library, as most candidates from the primary screen are not advanced to the confirmatory stage. A third possibility is the presence of dose response activity but lack of single dose for a set of compounds, which mostly occurs for hand-selected compounds believed to possibly be active. This scenario is rarely encountered in the public domain, and usually with an extremely small number of compounds. However, it is more common within industry, as detailed in our experiments (Section 3.6). The three scenarios are summarised below (Table 1).

Table 1. Summary of the different settings encountered when working with multi-fidelity HTS data. The presence and absence of a data type are symbolised by '✓', respectively '✗'. The fractions are representative for both the public and private AstraZeneca data.

Scenario	Single dose	Dose response	Fraction of data
Paired (SD, DR)	✓	✓	$< 0.01\%$
SD only	✓	✗	$> 99\%$
DR only	✗	✓	$< 0.001\%$

All molecular datasets are filtered for duplicates, stereoisomers, charged species, and large molecules (Supplementary Information 1). Overall, the same preprocessing steps are applied to the public and private AstraZeneca datasets, with any differences being presented in the following sections.

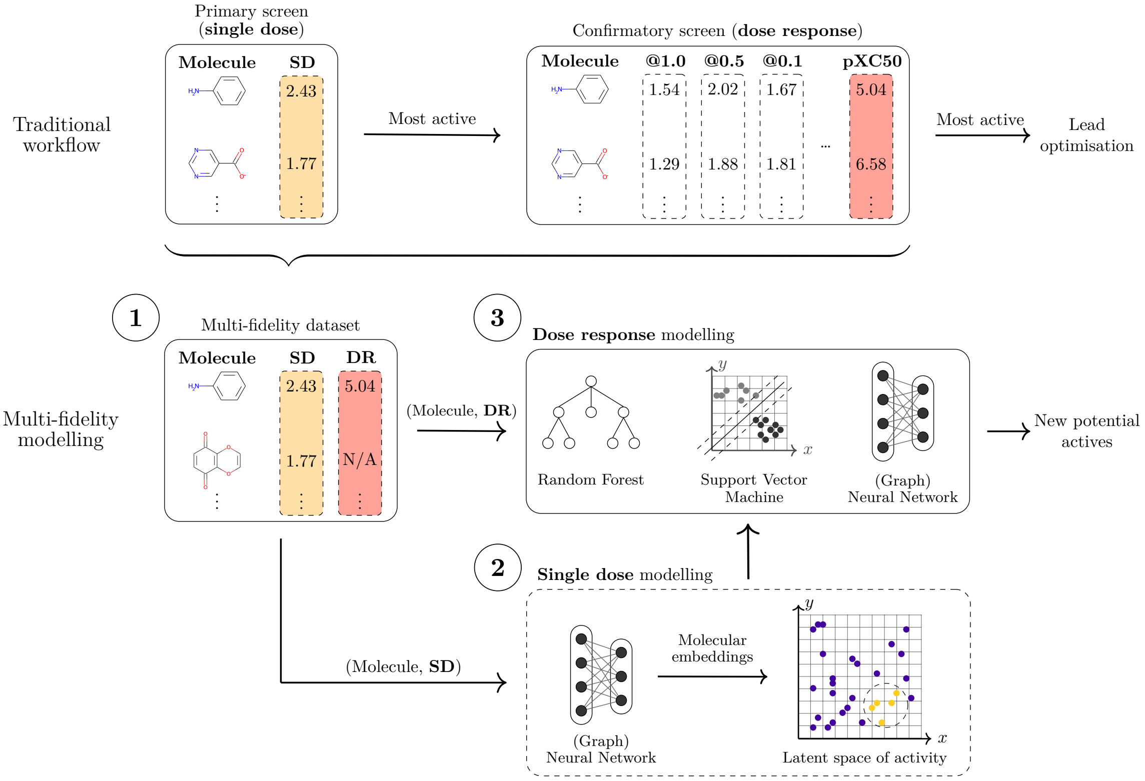


Figure 1. (Top) A traditional workflow consists of a massive but noisy primary screen on a large fraction of the available compound library, measuring the activity at a single concentration (single dose). Compounds that are regarded as active are manually selected for a confirmatory screen, which measures the activity at multiple different concentrations (dose response), usually summarised in a single number such as the pIC50 (the term 'pXC50' includes other variations). Finally, the most active compounds are manually selected for lead optimisation. (Middle, bottom) Our proposed multi-fidelity modelling framework illustrated with 3 high-level steps. Firstly, the corresponding single dose (SD) and dose response (DR) data are assembled into a multi-fidelity dataset. DR is only available for a fraction of the entire dataset, hence some compounds are not available ('N/A'). For step 2, a neural network is trained on the large primary screen, modelling a large and diverse chemical space of activity. Finally, at step 3, the molecular structure, supplemented with the representations learnt at step 2, is used train models of confirmatory activity. Compounds that are reported active can be further explored experimentally.

2.1.1 Public datasets

We manually searched PubChem using relevant keywords[1] and selected assays that have SD and DR measurements available. For some bioassays both modalities are available under the same assay identifier (AID); however, the majority are reported as separate confirmatory screening assays. It is possible that a single primary screening assay has multiple associated confirmatory assays, in which case we usually select and count the individual (SD AID, DR AID) dataset pairs that arise towards the total number of datasets. Primary screening information often includes replicate measurements, in which case they are averaged to provide a single value. For DR, with the exception of the fluorescence polarisation datasets, the readout values are transformed to the corresponding 'pXC50' unit. Overall, the collection shows a pronounced amount of heterogeneity in terms of assay type, screening technologies, concentrations, scoring metrics, protein targets, and scope, totalling 23 multi-fidelity datasets (each consists of paired SD and DR data).

We summarised essential details such as the SD and DR measurement types, the number of compounds in each screen, and the Pearson correlation coefficient for the existing paired (SD, DR) readouts for each of the 23 datasets that we used in the main analysis (Table 2). The selection criteria for the datasets included **(1)** diversity in the SD and DR types and sizes, as well as the assay format (biochemical, cell-based), **(2)** SD/DR correlation, and **(3)** availability of at least 20 compounds with dose-response readouts that are not associated with single-dose values. Each rule led to a different number of datasets: based on the diversity criterion, 8 datasets were selected (independent of Pearson's r), according to the

[1] Such as 'HTS', 'primary', 'confirmatory', etc.

Table 2. Summary of the 23 public multi-fidelity HTS datasets, including the PubChem AID, SD and DR measurement types, size of the datasets (denoted by #), the Pearson correlation coefficient (r) for the paired SD/DR measurements, and the associated p value. If the confirmatory data is available separately, both AID columns are populated, otherwise the SD dataset includes the DR data. The first 8 rows represent our starting set of public multi-fidelity data, the following 10 rows correspond to the datasets with the highest SD/DR correlation, and the last 5 rows summarise datasets that were added to support the transfer learning evaluation. The shortened words stand for: Inh., Inhibition; Act., Activation; Ind., Induction, FP, fluorescence polarisation.

DR AID	SD AID	SD type	DR type	# SD	# DR	r	p value
1259350	1224905	Z-score	FP	202,486	569	0.41	2.11×10^{-24}
1259418	1259416	Act.	pAC50	59,447	711	−0.37	1.97×10^{-24}
449756	435005	% change in signal	LogAC50	289,447	1,811	0.25	3.59×10^{-27}
–	449762	Inh. @25 µM	IC50	311,910	1,754	0.20	6.04×10^{-18}
–	1465	Fold Ind. @50 µM	EC50	205,193	980	−0.14	1.72×10^{-5}
1259375	1259374	Inh. @2.6 µM	LogIC50	614,427	348	0.10	6.89×10^{-2}
–	1949	Inh. @10 µg/mL	IC50 (µg/mL)	98,472	1,688	0.09	9.48×10^{-05}
1431	873	Inh. @5 µM	IC50	204,361	1,215	0.08	8.22×10^{-3}
–	504329	Inh. @12.5 µM	IC50	319,080	902	0.79	7.85×10^{-192}
–	1445	Inh. @30 µM	IC50	207,096	655	0.78	6.06×10^{-137}
624273	588549	Act. @12.48 µM	pAC50	337,483	359	0.70	1.55×10^{-54}
624326	602261	Act. @15 µM	IC50	343,811	985	0.68	1.03×10^{-133}
–	624330	Inh. @30 µM	IC50	324,979	1,570	0.66	2.30×10^{-198}
504941	488895	Act.	pAC50	321,242	161	0.63	4.15×10^{-19}
720512	652162	Act. @9.99 µM	pAC50	264,972	109	0.62	9.55×10^{-13}
624474	624304	Inh. @21.8 µM	IC50	345,553	1,327	0.58	1.43×10^{-121}
493155	485273	Inh. @20 µM	IC50	314,791	973	0.58	3.80×10^{-88}
435010	2221	Act.	LogEC50	280,006	1,797	0.56	1.27×10^{-149}
463203	2650	Act. @10 µM	LogAC50	300,560	721	0.42	1.83×10^{-31}
1259420	1259416	Act.	pAC50	59,447	174	−0.28	1.86×10^{-4}
2382	2098	Act. @7.5 µM	EC50	287,633	2,239	−0.24	1.29×10^{-29}
687027	652154	Act. @12.62 µM	pAC50	281,074	1,024	0.10	1.72×10^{-3}
504313	2732	Inh. @10 µM	IC50	208,123	855	−0.09	5.84×10^{-3}

second criterion, the top 10 datasets ranked by descending Pearson's r (absolute value) were selected, and finally based on the third criterion a total of 5 additional datasets qualified. The resulting heterogeneity of the 23 datasets allows the quantification of predictive performance effects based on properties such as the dataset size and the agreement between data types. For the 23 public datasets, the total number of unique primary interactions is 6,122,146, with 22,927 unique dose-response interactions.

2.1.2 AstraZeneca datasets

Similarly to the selection procedure for public assays described previously, historical AstraZeneca assays were manually searched and selected when satisfying conditions such as having more than 1 million compounds in the primary screen and multiple confirmatory screens available (usually referred to as dose response *rounds*). Based on these criteria, we selected 14 unique SD datasets, each associated with at least one DR dataset, for a total of 19 multi-fidelity datasets (Table 3). In this paper, each AstraZeneca dataset is assigned an arbitrary identifier such as AZ-SD1 for single dose, AZ-DR-R1 for DR (regression), and AZ-DR-C1 for DR (classification).

Differently from the public data, the emphasis is now on expanding the number of compounds screened in both single dose and dose response, generally surpassing that of the public repositories by a factor close to 6: the average number (± standard deviation) of SD compounds for the 23 public datasets is

268,765 ± 114,204, compared to 1,628,221 ± 247,472 for the 19 AstraZeneca datasets, and 997 ± 598 DR compounds for the public datasets compared to 5,419 ± 3,328 for the AstraZeneca datasets. The total number of unique primary interactions is 22,734,533, with 101,344 unique dose-response interactions.

The single primary screening scoring metric is the Z-Score, a normalisation method that represents the number of standard deviations from the population mean, with all but two of the dose response datasets using the pIC50, the others having only classification labels available. In the latter case, we adopted a conservative approach and labelled all inconclusive or irregular measurements as inactive, thus binarising the dataset (active or inactive). The raw HTS data was filtered using the same procedures described in Supplementary Information 1. Overall, we assembled a collection of 14 single dose datasets with 18 associated dose response data tables, with an additional dataset (AZ-DR-R4 1+2R – AZ-SD4) where we combined the DR measurements from two confirmatory screens (rounds) in a single multi-fidelity dataset to assess the influence of more data in the experiments, resulting in 19 SD/DR multi-fidelity datasets. The multi-fidelity datasets for the individual rounds are considered separate (AZ-DR-R4 1R – AZ-SD4, respectively AZ-DR-R4 2R – AZ-SD4). For the classification datasets, the point-biserial correlation coefficient was computed between the Z-Score and the activity label in place of the Pearson correlation coefficient.

Table 3. Summary of the AstraZeneca multi-fidelity datasets, including the SD and DR dataset names, SD and DR measurement types, size of the datasets (denoted by #), the Pearson correlation coefficient (r) for regression datasets or the point-biserial correlation coefficient (classification datasets) for the paired SD/DR measurements, and the associated p value. If for the same primary screening data there are multiple confirmatory screens available, each pair is represented through a different row in the table, where the DR name reflects the screening round (R) used. A p value of 0 indicates that the value is below the used machine precision, i.e. an extremely low value. Act., Activation.

DR name	SD name	SD type	DR type	# SD	# DR	r	p value
AZ-DR-R1	AZ-SD1	Z-Score	pIC50	1,700,745	6,522	−0.77	0
AZ-DR-R2	AZ-SD2			1,676,309	3,420	−0.72	0
AZ-DR-R3	AZ-SD3			1,970,086	9,654	−0.67	0
AZ-DR-R4 2R	AZ-SD4			1,370,897	914	−0.66	9.71×10^{-110}
AZ-DR-R5	AZ-SD3			1,970,086	9,523	−0.66	0
AZ-DR-R6	AZ-SD5			1,360,029	3,467	−0.66	0
AZ-DR-R2	AZ-SD6			1,013,581	11,828	−0.64	0
AZ-DR-R4 1+2R	AZ-SD4			1,370,897	1,615	−0.62	3.77×10^{-169}
AZ-DR-R4 1R	AZ-SD4			1,370,897	1,073	−0.58	1.14×10^{-68}
AZ-DR-R7	AZ-SD7			1,742,284	7,416	−0.53	0
AZ-DR-R8	AZ-SD7			1,742,284	6,909	−0.49	0
AZ-DR-R9	AZ-SD8			1,753,721	10,091	−0.46	0
AZ-DR-R10	AZ-SD9			1,581,928	399	−0.33	8.61×10^{-11}
AZ-DR-R11	AZ-SD10			1,671,471	4,488	−0.30	1.05×10^{-83}
AZ-DR-R12	AZ-SD11			1,747,502	5,642	−0.28	1.59×10^{-99}
AZ-DR-R13	AZ-SD11			1,747,502	4,698	−0.19	3.64×10^{-38}
AZ-DR-R14	AZ-SD12			1,962,638	6,511	−0.14	7.28×10^{-18}
AZ-SD13	AZ-DR-C1	Z-Score	Binary Act.	1,482,258	4,901	−0.22	3.50×10^{-52}
AZ-SD14	AZ-DR-C2			1,701,084	4,260	0.32	7.93×10^{-104}

2.2 Machine learning strategy

Multi-fidelity data modelling is performed in dose response space, by splitting the DR datasets into train (80%), validation (10%), and test (10%) sets, and incorporating SD data as described in the following sections. Each dataset is split five times based on different random seeds. The models are trained on all the resulting splits and the reported results aggregate the metrics from the corresponding models.

To validate our methodology, we evaluate three different classes of machine learning algorithms: random forests (RF), support vector machines (SVM), and graph neural networks (GNN). The validation set is used to guide the hyperparameter search for the RF and SVM models, and as part of an early stopping mechanism for the deep learning models. The final results are reported on the test set. Architectural details are provided (Supplementary Information 2), and all evaluated model configurations are listed (Supplementary Information 3, Supplementary Tables 1 to 3), amounting to just under 20,000 different models.

2.2.1 Shallow models

The RF and SVM algorithms provided by the open-source `scikit-learn` Python library can be used for both regression and classification tasks, requiring the input molecular structure to be pre-processed into a vector representation. We consider two different input representations: **(1)** Morgan fingerprints, computed using the open-source Python library RDKit with the function `GetMorganFingerprintAsBitVect()` and the parameters `radius=3` (increased from the default of 2), `nBits=2048` (default), and **(2)** a list of physical-chemical (PhysChem) descriptors, using the complete list provided by RDKit and computed with the function `MolecularDescriptorCalculator()`, totalling 208 PhysChem descriptors (as of March 2022).

In addition to the default `scikit-learn` hyperparameters for RF and SVM, we perform a hyperparameter search using `GridSearchCV` and negative mean squared error scoring to select the best models from a search space that balances coverage with reasonable training times (Supplementary Information 2). The best model configuration according to the hyperparameter search for each dataset was selected and used throughout the paper.

2.2.2 Deep learning models

We designed and implemented a novel deep learning architecture based on the variational graph autoencoder (VGAE). The VGAE is an unsupervised learning framework that exploits graph convolutional layers to learn directly from the non-Euclidean graph structure and the associated node (and possibly edge) features. As a member of the variational autoencoder family, the VGAE has an encoder that learns to compress the molecular information into a low-dimensional latent space and a decoder that reconstructs the original connectivity information. Here, the convolutions are used to learn and propagate atom-wise representations according to the connectivity imposed by the bonds, which are then aggregated into a single molecule-level representation or embedding (a fixed dimension vector) [Bro+21].

The resulting molecular representation can be further processed by a fully-connected neural network outputting the prediction in dose response space using an appropriate supervised loss, in an end-to-end fashion. Thus, the resulting model incorporates elements of unsupervised learning (compression-decompression) while being guided by the experimental readouts, an architecture we refer to as a ‘guided VGAE’. While this deep learning approach outputs bioactivity predictions, it also produces *molecular embeddings* that can be used in downstream analyses independent of the original emitting models, for example as additional input to both deep learning and classical methods such as RF and SVR.

Additional complexity is introduced by the choice of node (atom) aggregation functions that compute a compressed, molecule-level representation. The existing literature proposes simple permutation-invariant functions such as summation, mean, or maximum. In this work, we perform multi-fidelity integration experiments with three aggregation functions: summation, mean, and a new data-driven *neural aggregator* that we introduce in order to make better use of the multi-million scale information that is available. The neural aggregator is implemented as a fully-connected neural network that produces a molecule-level embedding from the concatenated atom (node) features of each graph. The deep learning and neural aggregator architectures are described in detail in Supplementary Information 2.

2.2.3 Experimental design

All the presented algorithms can be supplemented with the single dose data in different forms. Thus, we distinguish between the *base* and *augmented* machine learning models. Each of the tested machine learning models – RF, SVR, and the guided VGAE – can exist in the base or augmented forms. The base models are trained using only molecular structural information, i.e. either Morgan fingerprints or physical-chemical (PhysChem) descriptors for the shallow models (i.e. RF and SVR), or the molecular graphs for the deep learning models. The goal of predicting the dose-response bioactivity values (pXC50) or the confirmatory activity flag for classification datasets. This is in line with current industrial practices

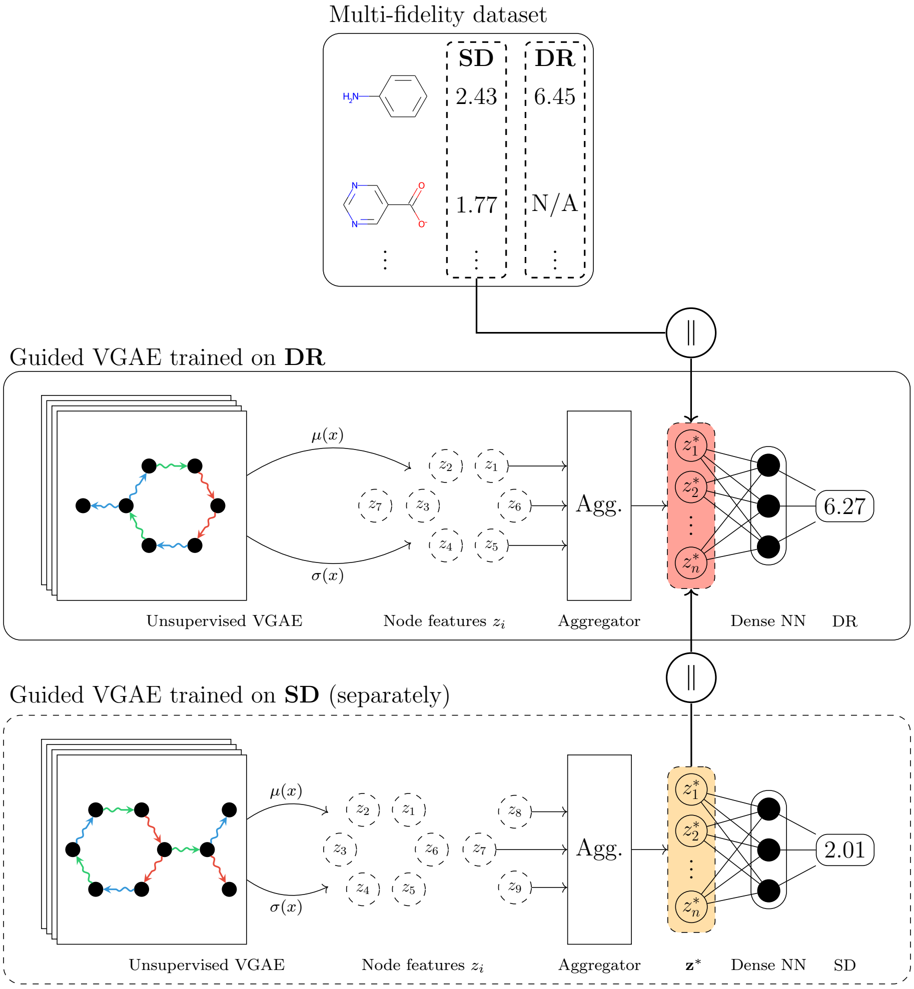


Figure 2. The proposed guided VGAE architecture presented diagrammatically. An **SD** model is first trained with supervised SD information, end-to-end, to produce graph (molecule) embeddings $\mathbf{z}^*$. A different model with the same architecture can then be trained to predict **DR** values, by concatenation with either the SD embeddings or the SD labels (only one at a time). The *Aggregator* is either global sum or mean pooling over the nodes, or a neural network. The symbol ‖ denotes concatenation. The models do not currently use bond (edge) features.

regarding high-throughput screening modelling, as only the highest-quality measurements (confirmatory) are used; however, this also limits the training set sizes to typically under 10,000 compounds, motivating the goal of integrating millions of related data points from primary screens. Consequently, the base models act as the reference point for each dataset.

We explore two different methods to incorporate SD data. To represent the 'best case' scenario for benefitting from SD observations, we append the SD label directly to the input fingerprint or the molecule-level embedding in the case of the VGAE (for the train, validation, and test sets). However, this necessitates that the SD label is available for each compound of interest, in a similar fashion to the existing HTSFP method.

To circumvent this limitation, we devise a further augmentation strategy. Its first step is to train a guided VGAE model exclusively on the entirety of the primary screening data for a fixed number of epochs. The trained models produce molecular embeddings (carrying the SD signal) that can be incorporated into a separate model trained and evaluated in DR space. This integration is achieved by concatenating the SD embedding with the fingerprint (RF, SVR) or internal representation (guided VGAE).

The main advantage behind the second augmentation strategy is the capability of the trained SD guided VGAE to produce molecular embeddings for arbitrary compounds, i.e. compounds lacking primary screening data for the particular protein target. We also emphasise that the learnt SD embeddings can be, in theory, integrated in the same way into any machine learning algorithm. However, whether the classical machine learning models can exploit the information produced by deep learning or not is a question that needs to be answered. As for the other augmentation, the SD embeddings are added to the train, validation, and test splits of each dataset.

A high-level representation of the deep learning pipeline and the two augmentations is illustrated in Figure 2. Overall, we evaluate three model configurations: base, augmented with SD labels, and augmented with SD embeddings, for each of the three machine learning frameworks we consider: random forests, support vector machines and the guided VGAE. For each VGAE model we tested three possible aggregators: sum, mean, and the novel neural aggregator. This translates to three SD deep learning models for each dataset, each producing SD molecular embeddings. Thus, each evaluated model (including RF, SVM, and the guided VGAE) has three variations for the SD embeddings augmentation. Similarly, the base and SD labels augmented deep learning models are also tested with all three aggregators. All the possible configurations are listed in the Supplementary Information (Supplementary Tables 1 to 3).

2.2.4 Reported metrics

For the majority of datasets, the prediction target is a ‘pXC50’ score. Thus, for all regression datasets we calculate the mean absolute error (MAE), the root mean squared error (RMSE), the maximum error, and the coefficient of determination (R^2) to measure the agreement between the real experimental values and the model predictions for the test sets.

Although all metrics are useful and can provide unique insights into the performance of the models, the coefficient of determination was recently argued to be more informative than other alternatives (such as the symmetric mean absolute percentage error – SMAPE) and more interpretable than metrics such as the MAE and RMSE [CWJ21].

For the two binary classification datasets, we report the AUROC (area under the receiver operating characteristic curve), a well understood and adopted classification metric, and the Matthews correlation coefficient (MCC), which was also recently argued to be more informative and truthful than the accuracy and F_1 score, by accurately summarising the confusion matrix information into a single number [CJ20].

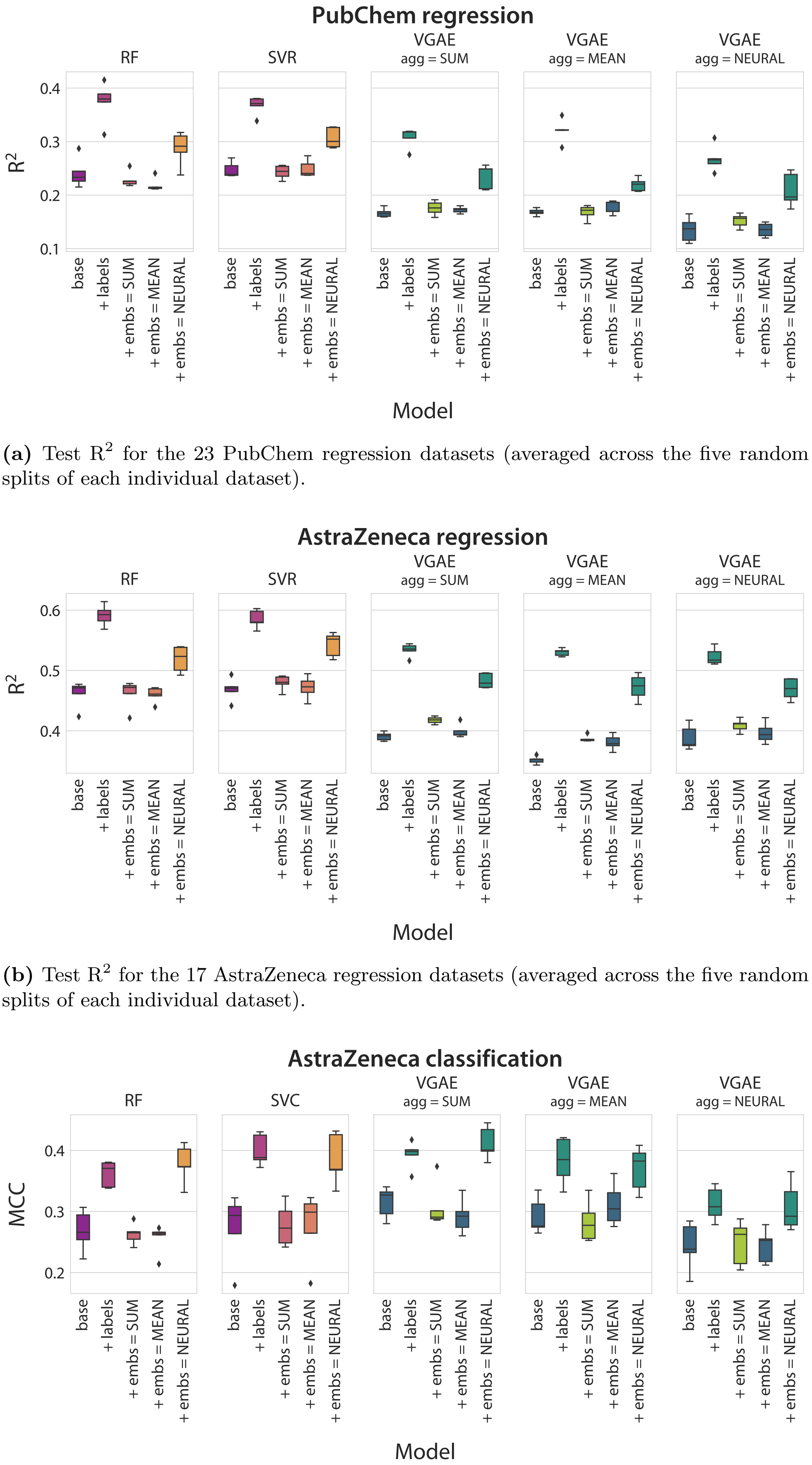
3 Results

To compare the predictive performance of the base and augmented models, we first report the aggregated test set results across three subgroups of datasets: **(1)** public (PubChem) regression, **(2)** AstraZeneca regression, and **(3)** AstraZeneca classification, using the R^2 metric for regression tasks and the MCC for classification (Figure 3, higher is better). Similar figures with MAE, RMSE, and maximum error for regression and the AUROC for classification are also provided (Supplementary Information 5 to Supplementary Information 7).

The general trends are followed by finer-grained, per-dataset summaries for a selection of AstraZeneca regression datasets (Figure 4) and PubChem regression datasets (Figure 5). Individual figures for each evaluated public and private dataset, with the same additional metrics as introduced previously are also available (Supplementary Information 8 to Supplementary Information 10). We discussed the statistical significance of certain dataset attributes such as the number of compounds in each dataset in Section 3.4, the trends towards more active or inactive confirmatory predictions for models integrating single dose data in Section 3.5, and the capability of the proposed methodology to generalise to compounds lacking single-dose measurements in Section 3.6.

3.1 Augmenting with experimental single dose readouts improves prediction performance

Starting the analysis with the SD labels augmentation (‘+ labels’ in Figure 3), firstly with the 23 public PubChem datasets, we observed an increase in performance above the base (non-SD augmented) models, on average, in R^2 from 0.241 to 0.374 for RF, from 0.247 to 0.367 for SVR, and from 0.167 to 0.307 for deep learning (using the sum aggregator). The results are coupled with decreases in MAE between 7% and 12% for the three ML algorithms (Supplementary Information 5). Large improvements are also observed for the AstraZeneca regression tasks, with an uplift in R^2, on average, from 0.461 to 0.591 for RF, from 0.469 to 0.585 for SVR, and from 0.391 to 0.534 for deep learning (sum aggregator). The corresponding improvements in MAE range from 15% to 18% (Supplementary Information 6).



(a) Test R^2 for the 23 PubChem regression datasets (averaged across the five random splits of each individual dataset).

(b) Test R^2 for the 17 AstraZeneca regression datasets (averaged across the five random splits of each individual dataset).

(c) Test MCC for the 2 AstraZeneca classification datasets (averaged across the five random splits of each individual dataset).

Figure 3. The predictive performance across the 3 multi-fidelity groups is reported on the test sets, using the R^2 for regression tasks and the MCC for classification tasks, combining metrics from each individual dataset including the five different random splits and presented using box plots with quartiles. The model configurations are summarised in Supplementary Tables 1 to 3. The panel titles denote the ML algorithm, including three different aggregation functions for the VGAE (‘SUM’, ‘MEAN’, ‘NEURAL’) and the x-axis labels (‘base’, ‘+ labels’ and ‘+ embs’) the augmentations.

For the two AstraZeneca classification datasets, the MCC improves, on average, from 0.269 to 0.362 for RF, from 0.273 to 0.400 for SVC, and from 0.315 to 0.393 for deep learning (sum aggregator), with uplifts in AUROC ranging from 6% to 11% (Supplementary Information 7).

Overall, the models augmented with primary screening data consistently outperform the baseline models. Out of the three dataset groups, the largest improvements are achieved for the public data, as the non-augmented performance is relatively low compared to the in-house AstraZeneca datasets, and the addition of single dose measurements can almost double the R^2.

The two shallow algorithms, RF and SVR, are almost evenly matched, with RF outperforming SVR for the two regression groups (PubChem and AstraZeneca), and SVC proving stronger on the AstraZeneca classification tasks.

The proposed deep learning architecture achieves the highest relative improvement ($\%R^2$) on the two regression groups (PubChem and AstraZeneca), although the performance of the base models is lower compared to RF and SVR. However, both the base and the SD labels augmented deep learning models outperform their shallow counterparts on the AstraZeneca classification tasks. Out of the three evaluated aggregation operators, the sum function proves to be consistently the best performing in the low-data regime of the base and SD labels augmented models (the models are trained only on the molecules with DR), with the mean and neural aggregators usually performing slightly worse, depending on the datasets.

3.2 Transfer learning with SD embeddings improves prediction performance

Continuing with the second augmentation strategy, we examine the three possible embedding types ('embs = SUM', 'embs = MEAN', and 'embs = NEURAL' in Figure 3), evaluated with each ML algorithm. To clarify, the SD and DR deep learning models are separate and can each use one of the three aggregation choices, resulting in 9 possible configurations. The three embedding types are also used to augment the RF and SVR models, corresponding to the last three entries in the 'RF' and 'SVR' panels in Figures 3a to 3c.

Firstly, we examined the training metrics for the SD deep learning models that produce the SD embeddings, after 150 epochs (Supplementary Table 4). When using neural aggregation, the training R^2 increases by up to 10 times on certain datasets, generally at least doubling the value of the simpler aggregators. Furthermore, we noticed a strong correlation ($r = 0.74, p = 6.2 \times 10^{-4}$) between the R^2 achieved after training on the SD models and the SD/DR correlation (Supplementary Figure 1).

We noticed an average increase in R^2 for the PubChem regression datasets from 0.241 to 0.287 for RF, from 0.247 to 0.307 for SVR, and from 0.167 to 0.235 for deep learning (sum aggregator for the DR model, neural aggregator for the SD model). The corresponding decreases in MAE range from 3% to 5% (Supplementary Information 5). For the AstraZeneca regression tasks, we noticed an increase, on average, in R^2 from 0.461 to 0.519 for RF, from 0.469 to 0.543 for SVR, and from 0.391 to 0.483 for deep learning (sum aggregator for DR, neural aggregator for SD). The improvements (decrease) in MAE are between 5% and 10% (Supplementary Information 6). For the collection of two AstraZeneca classification datasets, the MCC increased, on average, from 0.269 to 0.378 for RF, from 0.273 to 0.386 for SVC, and from 0.315 to 0.412 for deep learning (sum aggregator for DR, neural aggregator for SD). The AUROC registered increases in the range 8% − 10% (Supplementary Information 7).

The trends mirror those observed for the augmentation with single dose labels, namely consistent benefits in predictive performance compared to the base models across RF, SVM, and the guided VGAE models, although the effect is subtler for the majority of datasets. Importantly, we are able to show that the RF and SVM architectures can successfully incorporate and exploit the molecular embeddings emitted by separate deep learning models, a behaviour that was not immediately obvious as the three algorithms (RF, SVM, deep learning) are fundamentally different.

Extending our previous observations regarding the effect of node aggregator functions within graph neural networks, we can now conclude that the sum and mean operators are unable to capture the SD signal from the multi-million scale primary screens. The models augmented with these two types of molecular embedding generally perform the same, or even worse than the base models, across all evaluated ML algorithms and especially for the deep learning models. In contrast, the embeddings produced with the neural aggregator consistently outperformed the non-augmented models across the three ML strategies.

Interestingly, the models augmented with neural embeddings were the best performers on the AstraZeneca classification datasets (Figure 3c, including the shallow models and the guided VGAE with sum aggregation for the DR models), even compared to the SD labels augmentation. We also confirmed this on a number of regression datasets such as AZ-DR-R7 – AZ-SD7 (Figure 4a) and AZ-DR-R8 – AZ-SD7 (Figure 4b). However, for the majority of the regression datasets, the highest performance is achieved by the (augmented) shallow models.

3.3 Analysis of augmentation performance on a selection of datasets

To fully appreciate the benefits of multi-fidelity integration, we extend the analysis with a per-dataset discussion. In particular, we highlight results for datasets that exhibit differences from the general trends reported in the previous sections. To simplify the presentation, the configurations that produced poor results were omitted (i.e. sum and mean aggregation for SD, neural aggregation for DR).

3.3.1 AstraZeneca datasets

On datasets such as AZ-DR-R7 – AZ-SD7 (Figure 4a), AZ-DR-R8 – AZ-SD7 (Figure 4b), and AZ-DR-R9 – AZ-SD8 (Figure 4c), the best performing model in terms of R^2 is the SVR augmented with neural SD embeddings ('+ embs = NEURAL'). Other instances where the same configuration (SVR augmented with SD neural embeddings) outperforms the others were noticed, e.g. on AZ-DR-R11 – AZ-SD10 (Supplementary Figure 44d) and AZ-DR-R14 – AZ-SD12 (Supplementary Figure 38d).

The first three highlighted multi-fidelity datasets (Figures 4a to 4c) have a larger-than-average number of pre-split DR data points (Table 3), with 7,416, 6,909, and 10,091 respectively, and SD/DR correlations of $r = 0.53$, $r = 0.49$, and $r = 0.46$ respectively (absolute values), which are close to the average.

In contrast, AZ-DR-R2 – AZ-SD6 (Figure 4d) does not exhibit the same behaviour, despite being the dataset with the largest amount of DR compounds (11,828) and higher SD/DR correlation than the first three (0.64, absolute value). Still, the SD embeddings augmentation leads to improved R^2 compared to the base models. In this particular case, the SD dataset (AZ-SD6) has only 1,013,581 data points, significantly lower than the AstraZeneca average of 1,623,895 and representing the lowest in our collection. Furthermore, the three multi-fidelity datasets illustrated in Figures 4a to 4c all have over 1.7 million SD data points.

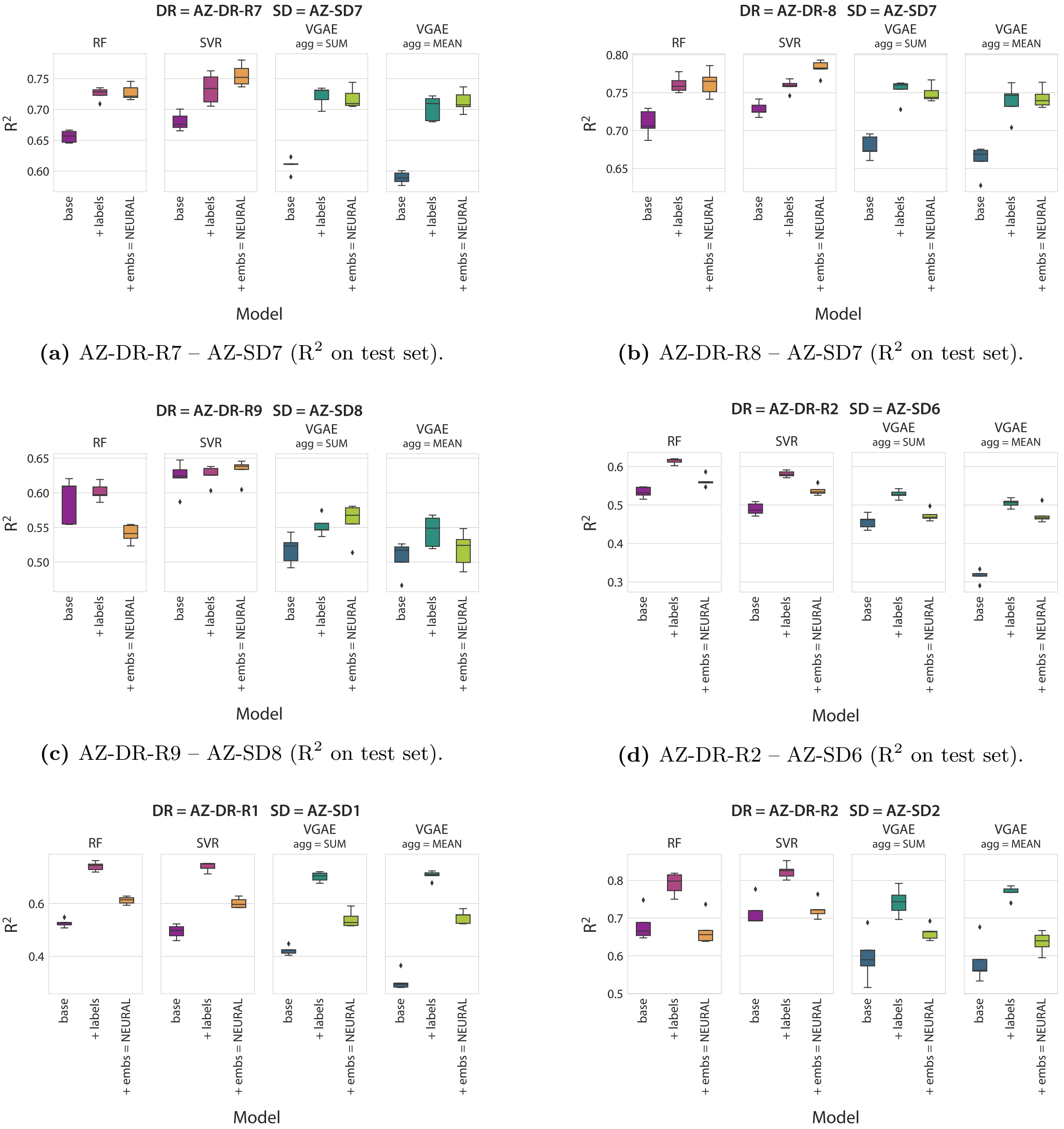
To offer a different perspective, we also visualised the two datasets with the highest SD/DR correlation, AZ-DR-R1 – AZ-SD1 (Figure 4e) and AZ-DR-R2 – AZ-SD2 (Figure 4f), with $r = 0.77$, respectively $r = 0.72$ (absolute value), and a DR compounds count of 6,522, respectively 3,420. For these two cases, the uplift in R^2 is among the largest for the SD labels augmentation compared to the base models. For AZ-DR-R1 – AZ-SD1, the achieved difference in R^2 between the SD labels augmentation and the base models (ΔR^2) for RF, SVR, and the guided VGAE models (sum aggregation) was of 0.216, 0.246, and 0.280, respectively. The corresponding decreases in MAE range between 29% and 32%. The augmentation with neural embeddings is beneficial, albeit to a lesser extent, with improvements in R^2 ranging from 16% to 29%, and decreases in MAE between 9% and 13%.

AZ-DR-R2 – AZ-SD2, the dataset with the second highest SD/DR correlation, sees smaller relative improvements. However, as one of the datasets with the stronger base performances (R^2 close to 0.7 for RF and SVR), it is remarkable to see improvements of the illustrated extent for the augmentation with SD labels ($R^2 > 0.8$ for SVR). AZ-DR-R3 – AZ-SD3, the third dataset in the SD/DR correlation hierarchy (Supplementary Figure 39), has slightly less than 10,000 DR compounds and behaves similarly to previously observed trends, with relative improvements for the SD labels augmentation between 29% and 58% in R^2, coupled with decreases in MAE ranging between 26% and 33%. In contrast, the two datasets with the lowest SD/DR correlation, AZ-DR-R14 – AZ-SD12 (Supplementary Figure 38d) and AZ-DR-R13 – AZ-SD11 (Supplementary Figure 37d), only see limited improvements for both augmentation strategies, generally under 10% in R^2, on average.

3.3.2 Public PubChem datasets

All machine learning models are challenged by the public datasets, achieving a baseline performance that is about two times lower in R^2, on average, than the AstraZeneca regression datasets (Figure 3).

As before, the study of individual cases shows that several datasets diverge from the trends described previously. One remarkable example is AID1445 (Figure 5a), the public dataset with the second-highest



(a) AZ-DR-R7 – AZ-SD7 (R^2 on test set).

(b) AZ-DR-R8 – AZ-SD7 (R^2 on test set).

(c) AZ-DR-R9 – AZ-SD8 (R^2 on test set).

(d) AZ-DR-R2 – AZ-SD6 (R^2 on test set).

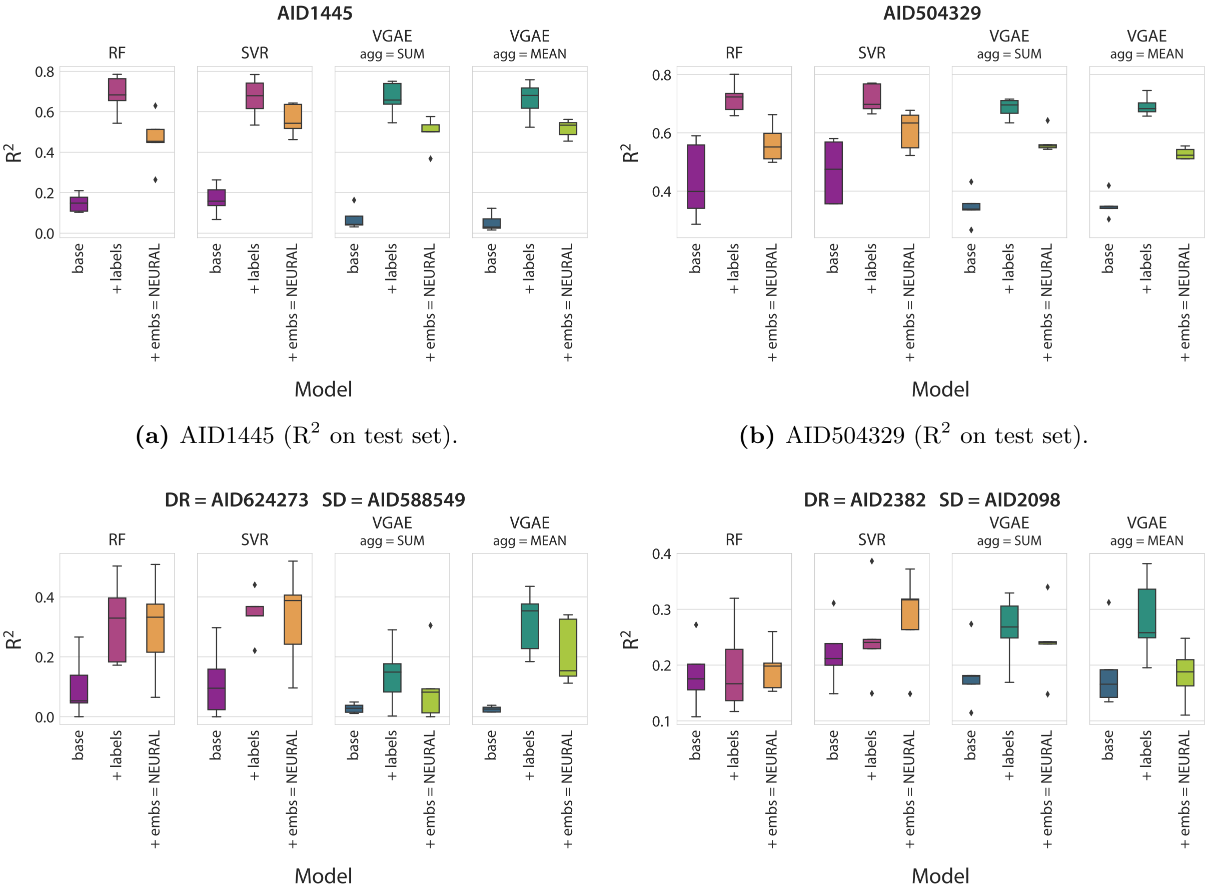
(e) AZ-DR-R1 – AZ-SD1 (R^2 on test set).

(f) AZ-DR-R2 – AZ-SD2 (R^2 on test set).

Figure 4. A small selection of AstraZeneca regression datasets, with the machine learning models limited to sum and mean aggregators for the DR models and neural aggregation for the SD models. The model configurations are summarised in Supplementary Tables 1 to 3. Each figure summarises results from the five different per-dataset random splits.

SD/DR correlation ($r = 0.78$) and where all base models performed poorly, with test R^2 scores of 0.149, 0.168, and 0.072 for RF, SVR, and deep learning, respectively. The SD labels augmentation increased the R^2, on average, to 0.686 for RF, to 0.671 for SVR, and to 0.666 for deep learning, with reductions in MAE, on average, between 40% and 44% (Supplementary Figure 8a). Augmenting with neural SD embeddings (sum aggregator for the DR models) was beneficial as well, the R^2 increasing, on average, to 0.462 for RF, to 0.560 for SVR, and to 0.497 for deep learning, the decreases in MAE ranging between 17% and 27%.

Notably, despite the base deep learning models underperforming compared to their shallow counterparts, both augmented guided VGAE models match the corresponding RF and SVR models. The public dataset with the highest correlation ($r = 0.79$), AID504329, also improves significantly, the augmentation with single dose labels leading to uplifts in R^2 between 65% and 98%, and MAE reductions in the range 28% – 35% (Supplementary Figure 18a), while the second augmentation strategy (same aggregator selection) produced increases in R^2 between 29% and 65%, with drops in MAE between 12% and 22%. Here, although the augmented performance is similar to AID1445, the base performance is considerably higher, resulting in smaller relative gains.



(a) AID1445 (R^2 on test set).

(b) AID504329 (R^2 on test set).

(c) AID624273 – AID588549 (R^2 on test set).

(d) AID2382 – AID2098 (R^2 on test set).

Figure 5. A small selection of PubChem regression datasets, with the machine learning models limited to sum and mean aggregators for the DR models and the neural aggregator for the SD models. The model configurations are summarised in Supplementary Tables 1 to 3. Each figure summarises results from the five different per-dataset random splits.

Similar but milder improvements can be observed for other datasets with high SD/DR correlation, such as AID624273 – AID588549 (Figure 5c) with $r = 0.70$. In addition, there are instances where augmenting with neural embeddings outperforms the SD labels augmentation even on the public datasets, such as AID2382 – AID2098 (Figure 5d) and AID1465 (Supplementary Figure 9).

Unfortunately, due to the limited amounts of data and low correlation for some public datasets, we encountered assays with poor model performance across the board, including the base and augmented configurations, such as AID1259375 – AID1259374, or AID1465 (Supplementary Figures 9 and 27). On the datasets with extremely low SD/DR correlation ($|r| < 0.10$), such as AID504313 – AID2732, AID1949, AID1431 – AID873, and AID687027 – AID652154 (Supplementary Figures 7, 10, 17 and 24), the augmentation strategies led to either minimal changes or slightly worse performance.

3.4 Effects of SD/DR agreement and dataset sizes on performance

Currently, there is no definitive answer to the question of the ‘right’ amount of data for high-throughput screening modelling. To investigate what factors affect the predictive performance of our models, we used multiple linear regression models to explain the evaluation metrics (e.g. the R^2) in terms of dataset attributes such as the SD/DR correlation, the number of SD molecules, and the number of DR molecules, as well as a categorical variable indicating the type of augmentation used. For the models trained on the public data, we observed statistically significant positive relationships between the R^2 and the SD/DR correlation and number of DR molecules ($p < 2 \times 10^{-16}$), as well as for the SD labels augmentation

($p < 2 \times 10^{-16}$) and the SD embeddings augmentation ($p = 1.31 \times 10^{-9}$), (Supplementary Table 5). Although we did not observe a significant contribution to the R^2 from the number of SD molecules, there are significant negative relationships between the other reported metrics (MAE, RMSE, maximum error) and the number of SD molecules, as well as the number of DR molecules (Supplementary Tables 6 to 8). In comparison, for the AstraZeneca datasets all encountered terms are significant, for the R^2 with $p < 2 \times 10^{-16}$ (Supplementary Table 9), and for the MAE and RMSE with $2 \times 10^{-16} < p < 10^{-8}$ (Supplementary Tables 10 and 11).

To relate dataset attributes specifically to the level of improvement induced by the inclusion of SD data versus the non-augmented base models, we performed a number of analyses on the $\Delta\mathrm{R}^2$ between both augmentations and the base models, for each machine learning algorithm. Firstly, we noticed that the agreement between the SD and DR datasets (Pearson's correlation coefficient) is strongly correlated with the $\Delta\mathrm{R}^2$ metric. For the models operating on public data and augmented with SD labels, we observed a correlation between the SD/DR correlation and the $\Delta\mathrm{R}^2$ of $r = 0.88$ for RF, of $r = 0.91$ for SVR, and of $r = 0.79$ for deep learning, with $p < 10^{-5}$ for each model (Supplementary Figures 49a to 49c). The models augmented with SD embeddings exhibited less pronounced correlation between the SD/DR correlation and the $\Delta\mathrm{R}^2$, with $r = 0.75,\ p = 4 \times 10^{-5}$ for RF, $r = 0.69,\ p = 2.5 \times 10^{-4}$ for SVR, and $r = 0.62,\ p = 0.00154$ for deep learning (Supplementary Figures 49d to 49f). For the AstraZeneca regression datasets, it was also possible to observe a statistically significant relationship between the two variables when augmenting with SD labels, with $r = 0.58,\ p = 0.01549$ for RF, $r = 0.55,\ p = 0.02259$ for SVR, and $r = 0.63,\ p = 0.00662$ for deep learning (Supplementary Figures 49g to 49i). The correlations for the SD embeddings augmentation were not statistically significant, although the guided VGAE model (Supplementary Figure 49l) indicates a possible relationship ($r = 0.43,\ p = 0.08373$).

In isolation, dataset attributes such as the number of SD molecules and the number of DR molecules are not coupled with the $\Delta\mathrm{R}^2$ (Supplementary Figures 50 and 51). However, multiple linear regression models with the $\Delta\mathrm{R}^2$ as the dependent variable show a positive statistically significant relationship for the number of DR molecules in the PubChem deep learning models augmented with SD labels ($p = 0.0179$, Supplementary Table 20), and a negative relationship in the AstraZeneca regression RF, SVR, and deep learning models augmented with SD labels ($p = 4.04 \times 10^{-4}$, Supplementary Table 22, $p = 4.43 \times 10^{-5}$, Supplementary Table 24, and $p = 0.0207$, Supplementary Table 26, respectively). Furthermore, the number of SD molecules is significant for the SD embeddings augmented deep learning models (positive relationship, $p = 0.0043$, Supplementary Table 27).

One particularly interesting example is given by two of the AstraZeneca multi-fidelity datasets, AZ-DR-R5 – AZ-SD3, with just under 2 million SD compounds, almost 10,000 DR compounds and an SD/DR correlation of $r = 0.66$, and AZ-DR-R2 – AZ-SD6, with just over 1 million SD compounds, almost 12,000 DR compounds, and an SD/DR correlation of $r = 0.64$ (Table 3). For these two datasets, the SD/DR correlations are very close and the number of DR compounds is also on the same scale, the only major difference being the number of SD compounds. For AZ-DR-R2 – AZ-SD6, only a modest increase in $\Delta\mathrm{R}^2$ is observed for the models augmented with SD embeddings, considering RF, SVR, and the guided VGAE, ($\Delta\mathrm{R}^2 < 0.05$), whereas for AZ-DR-R5 – AZ-SD3 the uplifts are several times higher, at $0.11 \leq \Delta\mathrm{R}^2 \leq 0.18$ (Supplementary Table 13). Furthermore, if we extend the analysis to AZ-DR-R6 – AZ-SD5, with almost 1.4 million SD compounds, 3.5K DR compounds, and an SD/DR correlation of $r = 0.66$, the corresponding $\Delta\mathrm{R}^2$ values lie between the two other datasets, although this comparison is not ideal as the number of DR compounds is lower than the other two. In contrast, for the three multi-fidelity datasets with the same SD set but different DR sets: AZ-DR-R4 1R - AZ-SD4, AZ-DR-R4 2R - AZ-SD4, AZ-DR-R4 1+2R - AZ-SD4 (Table 3), the differences in $\Delta\mathrm{R}^2$ are minimal (Supplementary Table 13).

3.5 Effect of the augmentation on confirmatory predictions

One of the most unique and valuable properties of the primary screening data is the vastness of the explored chemical space, especially when compared to the dose response experiments which are reserved for a fraction of the originally screened compound library. One aspect that was not explicitly investigated thus far is the direction taken by the augmented predictions: are the predictions shifted towards the more active or more inactive ends of the spectrum?

We decided to experiment with the historical AstraZeneca datasets, as they provide the largest amount of single dose interactions (> 1 million for every dataset). For each multi-fidelity AstraZeneca dataset, we separated the compounds with a Z-Score around 0 ($-0.5 \leq$ Z-Score ≤ 0.5) from the SD dataset, termed

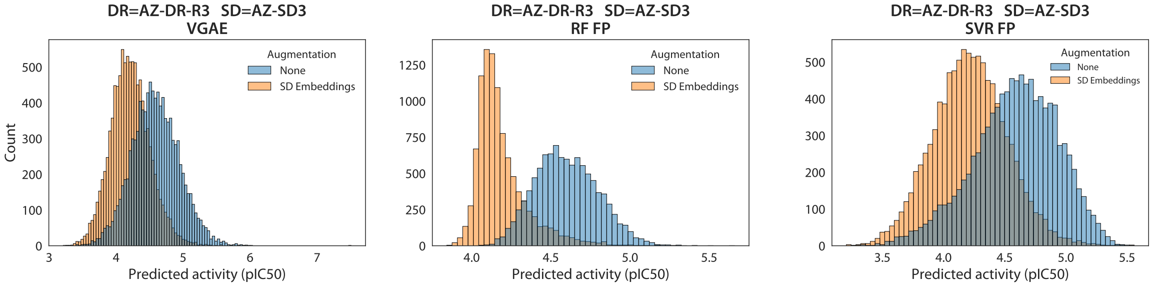


Figure 6. Histograms for the guided VGAE, RF, and SVR, comparing the SD embeddings augmentation to the base DR models models on a representative AstraZeneca multi-fidelity dataset (different scales on the y-axis).

'SD inactives', with 610,129 ± 80,370 compounds separated on average, and trained guided VGAE models on the rest of the SD data points (by extension called 'SD actives') using exactly the same methodology as presented previously. The trained 'SD actives' models were used to generate molecular embeddings for both the 'SD actives' and 'SD inactives' sets, noting that the 'SD inactives' were not seen during training; in other words, every compound from the original SD dataset was associated with an SD molecular embedding.

We repeated the evaluation procedure presented in Section 2.2.3 with the newly assembled data, focusing only on comparing the SD embeddings augmentation strategy to the base (non-augmented) models. The same five random splits of the DR data were used to train and validate DR models based on the guided VGAE, RF, and SVR algorithms, with the important difference that the existing 10% test set was replaced with a set of 10,000 diverse compounds selected from the 'SD inactives' sets for each dataset (*Diverse compounds selection*, Supplementary Information 1). As these new sets do not contain compounds with dose response measurements, it is only possible to compare the predicted pIC50 between the base and augmented models. The SD embeddings augmentation shifts the distribution of the predictions towards being inactive for the majority of datasets (Figure 6 and Supplementary Information 16). Furthermore, the SD embeddings have a pronounced effect on the predictions despite being produced by models that did not encounter the 10,000 molecules during training.

However, for a small number of datasets a few notable cases occurred, where the augmented models (guided VGAE, and RF, SVR with FP and PC as the input representation) predicted higher pIC50 than the base models; we report a selection of the top differences (Supplementary Table 28). A first observation is that the shallow models are generally more conservative in their predictions compared to the deep learning counterparts, rarely exceeding a pIC50 of 6.0. In contrast, the augmented deep learning models report high pIC50 values more often. However, on a non-trivial number of occasions the large gaps between augmented and base are mirrored for the shallow models, for example in AZ-DR-R9 – AZ-SD8, AZ-DR-R1 – AZ-SD1, and AZ-DR-R3 – AZ-SD3 to a lesser extent.

To help assess the quality of the predictions, we looked at the extreme pIC50 ranges for the models trained and evaluated in Section 3.2, more specifically low activity compounds with pIC50 < 3.5 and high activity compounds with pIC50 > 6.0, for each dataset where such compounds are available (**Supplementary File 2**). For these activity subsets, we computed the MAE based on the test set ground truth values (the R^2 is largely not statistically significant due to the low amount of samples). For the high activity subset, the augmented VGAE models had lower MAE than their shallow counterparts for 8 out of 14 datasets, and lower MAE than both the base and augmented shallow models for 7 out of 14 datasets. In contrast, for the low activity subset, the augmented VGAE models had lower MAE than the augmented RF and SVR models only for 4 out of 10 datasets, and lower MAE than both base and augmented shallow models for 3 out of 10 datasets.

More specifically, for AZ-DR-R2 – AZ-SD2 and AZ-DR-R9 – AZ-SD8 the top model for the high activity subset, according to the MAE on the test sets, is the base SVR FP, whereas for AZ-DR-R1 – AZ-SD1 and AZ-DR-R3 – AZ-SD3 the top model is the guided VGAE augmented with the SD embeddings. We propose two methods to validate high-activity predictions of this kind. Firstly, by examining compounds with high activity values reported by all models (deep learning, RF, and SVR), noting that the shallow models might not be always capable of fully exploiting the SD embeddings. Hence, high base activity values could also be considered. A few examples highlighted by this strategy are the 1st, 7th,

Table 4. Summary of the available multi-fidelity HTS datasets with compounds that have confirmatory screening information available but lack primary activity measurements, sorted by the number of such compounds, denoted by '# DR (no SD)'. For the private AstraZeneca collection, one dataset lacking primary data is evaluated in the context of 3 different trained DR models (AZ-DR-R4 – AZ-SD4), as reported before, whereas for AZ-DR-R2 – AZ-SD6 two different DR datasets lacking SD data are available.

Type	DR dataset (with SD)	SD dataset	# DR (no SD)	# SD	# DR	SD/DR r
Public (AID)	1259418	1259416	116	59,447	711	−0.37
	687027	652154	106	281,074	1,024	0.10
	1259420	1259416	103	59,447	174	−0.28
	504313	2732	39	208,123	855	−0.09
	463203	2650	27	300,560	721	0.42
	2382	2098	21	287,633	2,239	−0.24
Private (AZ)	AZ-DR-R4 1R	AZ-SD4	988	1,370,897	1,073	−0.58
	AZ-DR-R4 2R	AZ-SD4	988	1,370,897	914	−0.66
	AZ-DR-R4 1+2R	AZ-SD4	988	1,370,897	1,615	−0.62
	AZ-DR-R2	AZ-SD6	411	1,013,581	11,828	−0.64
	AZ-DR-R2	AZ-SD6	201	1,013,581	11,828	−0.64

11th, and 12th molecules listed for AZ-DR-R2 – AZ-SD2, the 1st, 4th, 6th, and 7th compounds for AZ-DR-R9 – AZ-SD8 and the entries for AZ-DR-R1 – AZ-SD1 and AZ-DR-R3 – AZ-SD3 (Supplementary Table 28).

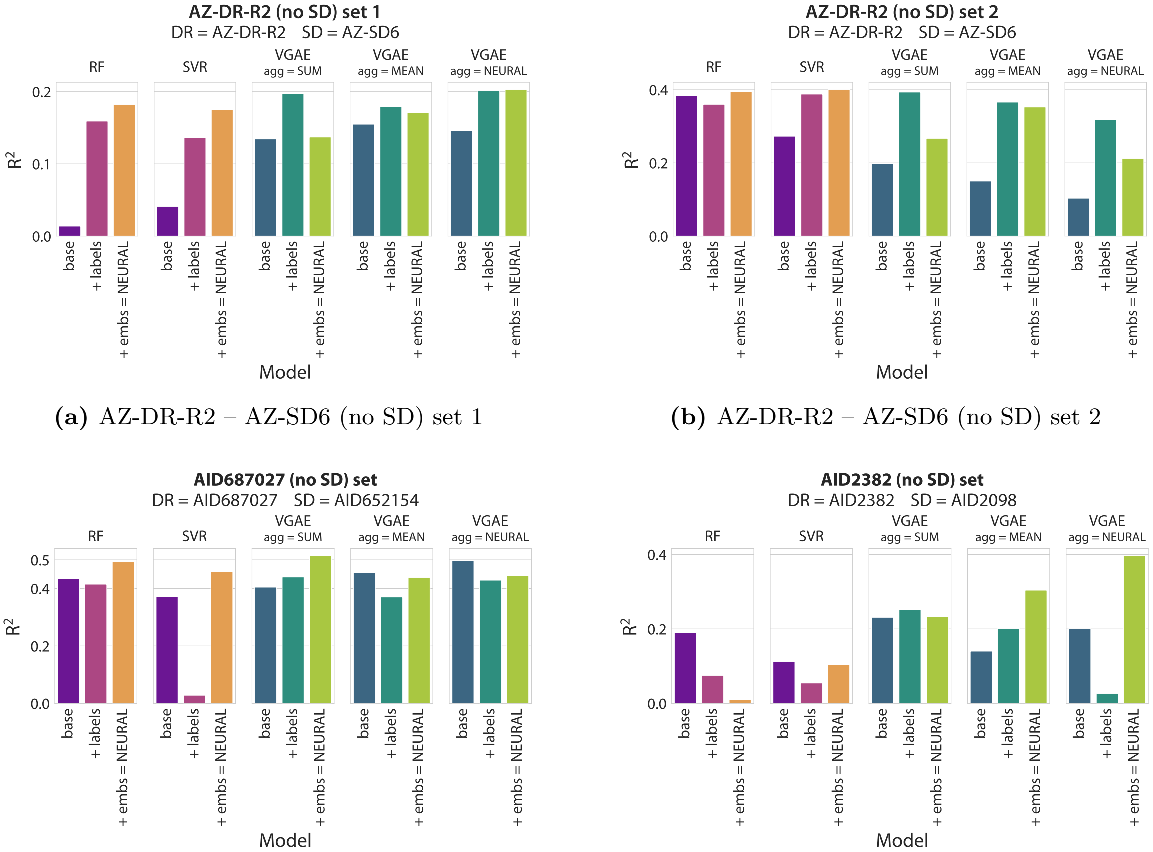
A different approach consists of relating the molecules suggested by deep learning with other screened compounds using similarity metrics such as the Tanimoto similarity. The compounds reported in this paper (Supplementary Table 28) are either not similar to other molecules in the primary screen (low Tanimoto similarity), or for the molecules with the highest similarity, which is still lower or equal to 0.8, the related compound is inactive according to the primary screen annotation. Likewise, similarity with compounds from the confirmatory screen is very low. The settings used for the similarity computation were Morgan fingerprints with 2048 bits and a radius of 3 in RDKit. As such, if any of the suggested molecules were experimentally validated as active, they would represent novel drug candidates that would have otherwise been missed.

3.6 Predictions for compounds lacking single-dose measurements

One of the most challenging tasks in the multi-fidelity context is predicting confirmatory-level activity for compounds lacking primary screening data, as this type of prediction relies on extrapolating single dose information from the trained SD models. A further complication is the sparsity of compounds where experimental confirmatory values are available but the corresponding single dose activity values are missing. From the entire search through PubChem, only 6 had more than 20 such compounds. Similarly, in the private AstraZeneca collection, only 3 DR sets lacking SD readouts are currently available, two being associated with the same SD dataset (Table 4).

In addition to the two augmentations presented so far, we also define a third possible augmentation, combining aspects from both strategies. As the experimental SD values are not available, we ask the trained SD models to produce single dose activity values, scalars that can be used with the existing implementation in exactly the same way as real SD readouts. To evaluate the SD embeddings and the (generated) SD labels on unseen molecules, the evaluation procedure was adjusted to train on the entirety of the DR data for a fixed number of epochs (200), as opposed to splitting into multiple train, validation, and test sets as before. Instead, the set of DR compounds lacking SD values was used as the test set.

Generally, we observe that both augmentations improve the predictive performance, as measured by the R^2, with a certain level of variability depending on the dataset. In the best case scenario, both augmentations led to a severalfold improvement for RF and SVR (AZ-DR-R2 – AZ-SD6 set 1, Figure 7a), as well as improvements for the guided VGAE models, and to doubling the performance of certain models on several datasets (AZ-DR-R2 – AZ-SD6 set 2 in Figure 7b, AID2382 – AID2098 in Figure 7d, AID1259420 – AID1259416 in Supplementary Figure 57f). Overall, the evaluation shows that the augmentations have a positive effect on performance, with a small number of instances with minimal, no, or negative effects (Supplementary Figures 56 and 57). Compared to the experimentally-derived SD values, the ones generated by machine learning appear more difficult to integrate successfully, in some cases even compared to the SD embeddings. This behaviour appears to be limited to the RF and SVR models, and more often for the PubChem data.



(a) AZ-DR-R2 – AZ-SD6 (no SD) set 1

(b) AZ-DR-R2 – AZ-SD6 (no SD) set 2

(c) AID687027 – AID652154 (no SD) set

(d) AID2382 – AID2098 (no SD) set

Figure 7. Evaluation results (R^2) for a selection of datasets that have experimentally derived DR values but lacking the associated SD activity. The '+ labels' augmentation makes use of SD labels generated by the trained SD models. The bold title refers to the set of data that was used for the evaluation (as a test set), whereas the secondary title refers to the DR and SD paired datasets that were used to train the ML models (no random splits).

4 Discussion

In this study, we proposed multi-fidelity HTS integration strategies built on a foundation of established and novel machine learning algorithms, with the goal of improving bioactivity predictions in confirmatory space. We demonstrated the benefits of leveraging primary screening data on a diverse selection of public and private datasets, with test set prediction improvements ranging between 25% and 85%, on average, depending on the specific algorithm and augmentation. On particular datasets, the integration of primary screening data enabled up to ×5 uplifts in predictive performance. Overall, we delivered strategies that can be applied to existing and upcoming HTS campaigns and that can generalise to unseen molecules, with the potential to highlight relevant compounds that would be missed by existing methods.

From a high-level perspective, we first validated two related hypotheses: (1) the primary screening data can be integrated into various machine learning algorithms and (2) the primary screening data has a positive effect on the predictive performance in confirmatory space.

4.1 Multi-fidelity integration with machine learning

To validate the two premises, we proposed two different augmentation strategies which enable the use of primary data in a fixed-dimension vector representation, a convenient format for most machine learning approaches, including classical and deep learning.

Directly including the SD measurements (first augmentation) led to considerable gains on the majority of datasets. The second augmentation strategy, which integrates molecular embeddings learnt by deep learning models, led to similar, but generally slightly lower improvements. However, we encountered several datasets where the SD embeddings augmentation proved to be the most successful in improving performance. This suggests that the learnt embeddings have the potential to be more useful than the SD values themselves for certain datasets.

Furthermore, we were able to establish that neural aggregation is the most effective option when a large amount of data is available, with the standard operators (such as sum or mean) still being preferable in low-data regimes. For HTS applications, this translates to neural aggregation when training on SD data, and one of the simpler functions when training on DR data.

On a practical note, we delivered techniques addressing two realistic HTS modelling scenarios: **(1)** improving confirmatory predictions when single dose measurements are available, and **(2)** generating high-quality predictions for new compounds that were not part of primary or confirmatory screens.

Regarding point **(1)**, we concluded that augmenting the DR models with experimental SD readouts is largely beneficial. Since primary screens are often performed on a multi-million scale, this translates to better confirmatory predictions for millions of compounds that would be too expensive to screen in the high-quality confirmatory assays. Furthermore, we suggest that in future HTS campaigns, all compounds of interest, for example from manual selection procedures or for historical reasons, are included in the primary screen. This step is likely to be inexpensive, but is expected to contribute to the prediction quality.

Point **(2)** refers to an even more challenging scenario, where the molecules in question do not possess experimentally-determined bioactivity values. Thus, predictions are generated by previously trained models on related SD and DR data. Although the augmentation strategies did not universally increase performance, we did encounter significant gains on several datasets, including two-fold uplifts. While this scenario remains less explored due to the lower amount of appropriate evaluation data, the demonstrated promise of the augmentations should motivate real-world use and the development of more comprehensive HTS assays and benchmarks.

4.2 Impact of the data size and quality

To explain the variability seen in the amount of performance gained by multi-fidelity integration, we sought to link the reported metrics to several dataset characteristics, which included SD/DR correlation, the size of the SD dataset, and the size of the DR dataset. In this process, we aimed to answer classic questions surrounding the computational chemistry space since the rise of Big Data, more specifically if more data is (always) helpful.

We first remarked that the uplift granted by the SD labels augmentation increases almost linearly with the agreement between SD and DR for the public PubChem datasets. The effect was strong enough to be observable even in isolation (without including other dataset attributes), and suggests that the quality of the measurements, more specifically in the primary screen, is the deciding factor for successful computational modelling of multi-fidelity HTS data. The relationship is less pronounced for the AstraZeneca data, likely due to the larger role played by the amount of data for both modalities.

Furthermore, for the AstraZeneca regression datasets (all larger than 1 million data points) we examined the metrics (R^2) after fully training the SD models. We noticed a strong positive relationship (Pearson's correlation coefficient) between the training set R^2 and the aforementioned SD/DR correlation between the two data modalities. This relationship suggests that the SD datasets that are not in agreement with the DR counterpart are noisy, reflected in the challenge posed to the deep learning models. In contrast, the SD datasets that are highly correlated with the DR measurements are easier to model and lead to more informative and helpful embeddings.

Our analysis revealed that the number of SD compounds and the number of DR compounds were statistically significant in explaining the reported performance metrics, as well as the uplifts (ΔR^2). Larger primary screens were significantly associated with the uplifts seen for the deep learning models using the embeddings augmentation on AstraZeneca data. In a case by case analysis looking at embeddings-augmented models on AstraZeneca datasets with similar profiles but vastly different number of SD compounds, we observed that the largest uplifts correspond to the largest primary screens. For this comparison, all dataset attributes except the number of SD compounds are very similar and could be considered fixed, leading us to theorise that the large difference in the primary screen size explains the uplifts.

Similar situations can be encountered for public datasets, for example AID493155 – AID485273, AID624474 – AID624304, and AID435010 – AID2221 (Supplementary Table 15). The largest uplifts for the models augmented with SD embeddings are seen on the dataset with the largest number of SD compounds out of the three. Additionally, the models augmented with SD labels for the three highlighted datasets, which do not depend on the number of SD compounds, perform very closely, especially for RF and SVR. This indicates that the number of SD molecules is an important factor for the models augmented with SD embeddings.

The size of the confirmatory screen was also significant in explaining the uplifts for certain RF, SVR, and deep learning models on the public and especially AstraZeneca data. In particular, having more DR data is associated with lower uplifts on the AstraZeneca datasets, and with higher uplifts for the deep learning models on PubChem datasets. One possible explanation for the former is that models with more DR training data have less space to improve since they already cover a larger chemical space. Another effect to take into consideration is the train/validation/test split ratio, which is kept constant for all evaluated datasets. It may be more difficult to achieve uplifts on the largest DR test sets (more than ten times larger than the smallest, for example AZ-DR-R10 – AZ-SD9 with 399 pre-split DR compounds versus AZ-DR-R2 – AZ-SD6 with 11,828), as they are more likely to contain compound classes and structures not seen during training.

Overall, these observations indicate that the deep learning models are more sensitive to the amount of data in the primary and confirmatory screens, possibly coupled with making better use of previously learnt molecular embeddings compared to RF and SVR. Generally, high amounts of SD and DR data, as well as good SD/DR correlation are required for the effective integration of SD data using learnt embeddings. For the public data, the relatively small size of the SD datasets and the low variability in the number of SD compounds between public datasets may considerably reduce the observable effect of more data.

4.3 Effects of incorporating a larger chemical space

For the majority of HTS experiments we expect only a very limited amount of molecules to truly interact favourably with the protein target. As successfully integrating the primary screening information should provide a global overview of the tested chemical space, we argue that the augmented models will produce more conservative bioactivity predictions, leading to more predicted inactives on the whole.

Indeed, the models augmented with the SD embeddings led to lower pIC50 predictions for the majority of AstraZeneca datasets, thus lowering the risk of reporting false positives. For a few datasets, we noticed that the augmented models also reported high activity for a small number of molecules with low base activity scores. This behaviour was mostly observed for the deep learning models. In our analysis, we established that for the most active predictions the deep learning models are at least as accurate as RF or SVR, while the classical algorithms are better suited for low-activity predictions. This is encouraging as deep learning predictions could translate to new promising compounds that would not be discovered by classical algorithms.

For AZ-DR-R2 – AZ-SD2, our analysis led to 39 compounds with largely increased predicted activity from the diverse set of 10,000 compounds (12 reproduced in Supplementary Table 28). Extrapolating these trends to an entire primary screen of 1 to 2 million compounds indicates that our method allows the selection of a few hundred or thousand molecules for a supplementary confirmatory round. Based on these observations, we hypothesise that using the suggested filtering strategies as a cost-effective selection step is a beneficial addition to live HTS assays, providing possible leads from compounds which otherwise might have been missed by manual inspection or existing techniques.

References

[Gib85] Alan Gibbons. *Algorithmic graph theory*. Cambridge Cambridgeshire New York: Cambridge University Press, 1985. ISBN: 9780521288811.

[BGV92] Bernhard E. Boser, Isabelle M. Guyon, and Vladimir N. Vapnik. 'A Training Algorithm for Optimal Margin Classifiers'. In: *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*. COLT '92. Pittsburgh, Pennsylvania, USA: Association for Computing Machinery, 1992, 144–152. ISBN: 089791497X. DOI: 10.1145/130385.130401. URL: https://doi.org/10.1145/130385.130401.

[Bre01] Leo Breiman. 'Random Forests'. In: *Machine Learning* 45.1 (Oct. 2001), pp. 5–32. ISSN: 1573-0565. DOI: 10.1023/A:1010933404324. URL: https://doi.org/10.1023/A:1010933404324.

[Bur+01] R. Burbidge et al. 'Drug design by machine learning: support vector machines for pharmaceutical data analysis'. In: *Comput Chem* 26.1 (Dec. 2001), pp. 5–14.

[TH03] Matthew W. B. Trotter and Sean B. Holden. 'Support Vector Machines for ADME Property Classification'. In: *QSAR & Combinatorial Science* 22.5 (2003), pp. 533–548. DOI: https://doi.org/10.1002/qsar.200310006. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/qsar.200310006. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/qsar.200310006.

[Sve+04] Vladimir Svetnik et al. 'Application of Breiman's Random Forest to Modeling Structure-Activity Relationships of Pharmaceutical Molecules'. In: *Multiple Classifier Systems*. Ed. by Fabio Roli, Josef Kittler, and Terry Windeatt. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 334–343. ISBN: 978-3-540-25966-4.

[PW07] D. A. Pereira and J. A. Williams. 'Origin and evolution of high throughput screening'. In: *Br J Pharmacol* 152.1 (Sept. 2007), pp. 53–61.

[Per10] Emanuele Perola. 'An Analysis of the Binding Efficiencies of Drugs and Their Leads in Successful Drug Discovery Programs'. In: *Journal of Medicinal Chemistry* 53.7 (Apr. 2010), pp. 2986–2997. ISSN: 0022-2623. DOI: 10.1021/jm100118x. URL: https://doi.org/10.1021/jm100118x.

[Mac+11] Ricardo Macarron et al. 'Impact of high-throughput screening in biomedical research'. In: *Nature Reviews Drug Discovery* 10.3 (Mar. 2011), pp. 188–195. ISSN: 1474-1784. DOI: 10.1038/nrd3368. URL: https://doi.org/10.1038/nrd3368.

[Pet+12] Paula M. Petrone et al. 'Rethinking Molecular Similarity: Comparing Compounds on the Basis of Biological Activity'. In: *ACS Chemical Biology* 7.8 (2012). PMID: 22594495, pp. 1399–1409. DOI: 10.1021/cb3001028. eprint: https://doi.org/10.1021/cb3001028. URL: https://doi.org/10.1021/cb3001028.

[Duv+15] David K Duvenaud et al. 'Convolutional Networks on Graphs for Learning Molecular Fingerprints'. In: *Advances in Neural Information Processing Systems*. Ed. by C. Cortes et al. Vol. 28. Curran Associates, Inc., 2015. URL: https://proceedings.neurips.cc/paper/2015/file/f9be311e65d81a9ad8150a60844bb94c-Paper.pdf.

[IS15] Sergey Ioffe and Christian Szegedy. 'Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift'. In: *Proceedings of the 32nd International Conference on Machine Learning*. Ed. by Francis Bach and David Blei. Vol. 37. Proceedings of Machine Learning Research. Lille, France: PMLR, July 2015, pp. 448–456. URL: https://proceedings.mlr.press/v37/ioffe15.html.

[Hel+16] Kazi Yasin Helal et al. 'Public Domain HTS Fingerprints: Design and Evaluation of Compound Bioactivity Profiles from PubChem's Bioassay Repository'. In: *Journal of Chemical Information and Modeling* 56.2 (2016). PMID: 26898267, pp. 390–398. DOI: 10.1021/acs.jcim.5b00498. eprint: https://doi.org/10.1021/acs.jcim.5b00498. URL: https://doi.org/10.1021/acs.jcim.5b00498.

[KW16] Thomas N. Kipf and Max Welling. *Variational Graph Auto-Encoders*. 2016. arXiv: 1611.07308 [stat.ML].

[AT+17] Han Altae-Tran et al. 'Low Data Drug Discovery with One-Shot Learning'. In: *ACS Central Science* 3.4 (2017). PMID: 28470045, pp. 283–293. DOI: 10.1021/acscentsci.6b00367. eprint: https://doi.org/10.1021/acscentsci.6b00367. URL: https://doi.org/10.1021/acscentsci.6b00367.

[KW17] Thomas N. Kipf and Max Welling. *Semi-Supervised Classification with Graph Convolutional Networks*. 2017. arXiv: 1609.02907 [cs.LG].

[Mof+17] John G. Moffat et al. 'Opportunities and challenges in phenotypic drug discovery: an industry perspective'. In: *Nature Reviews Drug Discovery* 16.8 (Aug. 2017), pp. 531–543. ISSN: 1474-1784. DOI: 10.1038/nrd.2017.111. URL: https://doi.org/10.1038/nrd.2017.111.

[Sch+17] Kristof Schütt et al. 'SchNet: A continuous-filter convolutional neural network for modeling quantum interactions'. In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper/2017/file/303ed4c69846ab36c2904d3ba8573050-Paper.pdf.

[BB18] Dean G. Brown and Jonas Boström. 'Where Do Recent Small Molecule Clinical Development Candidates Come From?' In: *Journal of Medicinal Chemistry* 61.21 (2018). PMID: 29920198, pp. 9442–9468. DOI: 10.1021/acs.jmedchem.8b00675. eprint: https://doi.org/10.1021/acs.jmedchem.8b00675. URL: https://doi.org/10.1021/acs.jmedchem.8b00675.

[JBJ18] Wengong Jin, Regina Barzilay, and Tommi S. Jaakkola. 'Junction Tree Variational Autoencoder for Molecular Graph Generation'. In: *CoRR* abs/1802.04364 (2018). arXiv: 1802.04364. URL: http://arxiv.org/abs/1802.04364.

[Mor+18] Paul Morgan et al. 'Impact of a five-dimensional framework on R&D productivity at AstraZeneca'. In: *Nature Reviews Drug Discovery* 17.3 (Mar. 2018), pp. 167–181. ISSN: 1474-1784. DOI: 10.1038/nrd.2017.244. URL: https://doi.org/10.1038/nrd.2017.244.

[Wu+18] Zhenqin Wu et al. 'MoleculeNet: a benchmark for molecular machine learning'. In: *Chem. Sci.* 9 (2 2018), pp. 513–530. DOI: 10.1039/C7SC02664A. URL: http://dx.doi.org/10.1039/C7SC02664A.

[Che+19] Deli Chen et al. 'Measuring and Relieving the Over-smoothing Problem for Graph Neural Networks from the Topological View'. In: *CoRR* abs/1909.03211 (2019). arXiv: 1909.03211. URL: http://arxiv.org/abs/1909.03211.

[Lau+19] Oliver Laufkötter et al. 'Combining structural and bioactivity-based fingerprints improves prediction performance and scaffold hopping capability'. In: *Journal of Cheminformatics* 11.1 (Aug. 2019), p. 54. ISSN: 1758-2946. DOI: 10.1186/s13321-019-0376-1. URL: https://doi.org/10.1186/s13321-019-0376-1.

[Stu+19] Noé Sturm et al. 'Application of Bioactivity Profile-Based Fingerprints for Building Machine Learning Models'. In: *Journal of Chemical Information and Modeling* 59.3 (2019). PMID: 30408959, pp. 962–972. DOI: 10.1021/acs.jcim.8b00550. eprint: https://doi.org/10.1021/acs.jcim.8b00550. URL: https://doi.org/10.1021/acs.jcim.8b00550.

[Vol+19] Dmitriy M. Volochnyuk et al. 'Evolution of commercially available compounds for HTS'. In: *Drug Discovery Today* 24.2 (2019), pp. 390–402. ISSN: 1359-6446. DOI: https://doi.org/10.1016/j.drudis.2018.10.016. URL: https://www.sciencedirect.com/science/article/pii/S1359644618302423.

[Yan+19] Kevin Yang et al. 'Analyzing Learned Molecular Representations for Property Prediction'. In: *Journal of Chemical Information and Modeling* 59.8 (2019). PMID: 31361484, pp. 3370–3388. DOI: 10.1021/acs.jcim.9b00237. eprint: https://doi.org/10.1021/acs.jcim.9b00237. URL: https://doi.org/10.1021/acs.jcim.9b00237.

[CJ20] Davide Chicco and Giuseppe Jurman. 'The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation'. In: *BMC Genomics* 21.1 (Jan. 2020), p. 6. ISSN: 1471-2164. DOI: 10.1186/s12864-019-6413-7. URL: https://doi.org/10.1186/s12864-019-6413-7.

[Cla20] David E. Clark. 'Virtual Screening: Is Bigger Always Better? Or Can Small Be Beautiful?' In: *Journal of Chemical Information and Modeling* 60.9 (2020). PMID: 32463232, pp. 4120–4123. DOI: 10.1021/acs.jcim.0c00101. eprint: https://doi.org/10.1021/acs.jcim.0c00101. URL: https://doi.org/10.1021/acs.jcim.0c00101.

[Cor+20] Gabriele Corso et al. 'Principal Neighbourhood Aggregation for Graph Nets'. In: *CoRR* abs/2004.05718 (2020). arXiv: 2004.05718. URL: https://arxiv.org/abs/2004.05718.

[Gur+20] Oleksandr Gurbych et al. 'High throughput screening with machine learning'. In: *CoRR* abs/2012.08275 (2020). arXiv: 2012.08275. URL: https://arxiv.org/abs/2012.08275.

[Gö+20] Andreas H. Göller et al. 'Bayer's in silico ADMET platform: a journey of machine learning over the past two decades'. In: *Drug Discovery Today* 25.9 (2020), pp. 1702–1709. ISSN: 1359-6446. DOI: https://doi.org/10.1016/j.drudis.2020.07.001. URL: https://www.sciencedirect.com/science/article/pii/S1359644620302609.

[KSL20] Beomchang Kang, Chaok Seok, and Juyong Lee. ‘Prediction of Molecular Electronic Transitions Using Random Forests’. In: *Journal of Chemical Information and Modeling* 60.12 (2020). PMID: 33090804, pp. 5984–5994. DOI: 10.1021/acs.jcim.0c00698. eprint: https://doi.org/10.1021/acs.jcim.0c00698. URL: https://doi.org/10.1021/acs.jcim.0c00698.

[Sto+20] Jonathan M. Stokes et al. ‘A Deep Learning Approach to Antibiotic Discovery’. In: *Cell* 180.4 (2020), 688–702.e13. ISSN: 0092-8674. DOI: https://doi.org/10.1016/j.cell.2020.01.021. URL: https://www.sciencedirect.com/science/article/pii/S0092867420301021.

[Tow+20] Raphael J. L. Townshend et al. ‘ATOM3D: Tasks On Molecules in Three Dimensions’. In: *CoRR* abs/2012.04035 (2020). arXiv: 2012.04035. URL: https://arxiv.org/abs/2012.04035.

[TNJR20] Viet-Khoa Tran-Nguyen, Célien Jacquemard, and Didier Rognan. ‘LIT-PCBA: An Unbiased Data Set for Machine Learning and Virtual Screening’. In: *Journal of Chemical Information and Modeling* 60.9 (2020). PMID: 32282202, pp. 4263–4273. DOI: 10.1021/acs.jcim.0c00155. eprint: https://doi.org/10.1021/acs.jcim.0c00155. URL: https://doi.org/10.1021/acs.jcim.0c00155.

[Zho+20] Kaixiong Zhou et al. ‘Towards Deeper Graph Neural Networks with Differentiable Group Normalization’. In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., 2020, pp. 4917–4928. URL: https://proceedings.neurips.cc/paper/2020/file/33dd6dba1d56e826aac1cbf23cdcca87-Paper.pdf.

[Bro+21] Michael M. Bronstein et al. ‘Geometric Deep Learning: Grids, Groups, Graphs, Geodesics, and Gauges’. In: *CoRR* abs/2104.13478 (2021). arXiv: 2104.13478. URL: https://arxiv.org/abs/2104.13478.

[CWJ21] Davide Chicco, Matthijs J. Warrens, and Giuseppe Jurman. ‘The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation’. In: *PeerJ Computer Science* 7 (July 2021), e623. DOI: 10.7717/peerj-cs.623. URL: https://doi.org/10.7717/peerj-cs.623.

[Dre+21] Gabriel H. S. Dreiman et al. ‘Changing the HTS Paradigm: AI-Driven Iterative Screening for Hit Finding’. In: *SLAS DISCOVERY: Advancing the Science of Drug Discovery* 26.2 (2021). PMID: 32808550, pp. 257–262. DOI: 10.1177/2472555220949495. eprint: https://doi.org/10.1177/2472555220949495. URL: https://doi.org/10.1177/2472555220949495.

[Sta+21] Megan Stanley et al. ‘FS-Mol: A Few-Shot Learning Dataset of Molecules’. In: *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*. 2021. URL: https://openreview.net/forum?id=701FtuyLlAd.

[ASB] Stevan Aleksić, Daniel Seeliger, and J. B. Brown. ‘ADMET Predictability at Boehringer Ingelheim: State-of-the-Art, and Do Bigger Datasets or Algorithms Make a Difference?’ In: *Molecular Informatics* n/a.n/a (), p. 2100113. DOI: https://doi.org/10.1002/minf.202100113. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/minf.202100113. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/minf.202100113.