# *Practical considerations for active machine learning in drug discovery*

# Practical considerations for active machine learning in drug discovery

Daniel Reker[1,2,*]

[1] Koch Institute for Integrative Cancer Research and MIT-IBM Watson AI Lab, Massachusetts Institute of Technology, Cambridge MA, USA

[2] Division of Gastroenterology, Hepatology and Endoscopy, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston MA, USA

* Correspondence: reker@mit.edu

**Active machine learning enables the automated selection of the most valuable next experiments to improve predictive modelling and hasten active retrieval in drug discovery. Although a long established theoretical concept and introduced to drug discovery approximately 15 years ago, the deployment of active learning technology in the discovery pipelines across academia and industry remains slow. With the recent re-discovered enthusiasm for artificial intelligence as well as improved flexibility of laboratory automation, active learning is expected to surge and become a key technology for molecular optimizations. This review recapitulates key findings from previous active learning studies to highlight the challenges and opportunities of applying adaptive machine learning to drug discovery. Specifically, considerations regarding implementation, infrastructural integration, and expected benefits are discussed. By focusing on these practical aspects of active learning, this review aims at providing insights for scientists planning to implement active learning workflows in their discovery pipelines.**

Active machine learning is an active field of applied machine learning research, striving to conceive experimental selection functions that aid in identifying the most valuable next experiment [1] – thereby putting machine learning into the driver seat of iterative molecular design efforts (Figure 1). [2,3] To this end, novel experimental protocols are either generated by the algorithm or selected from a pre-generated set of possible experiments. [1] The value of an experiment can be defined by a range of different notions, but commonly includes some anticipated benefit to improve the performance of the machine learning method – most commonly derived from predictive confidence measures that enable selecting data with the highest predictive uncertainty. [2] Additional formalizations can be added that aid steering experimental design towards molecular structures with desired properties, such as improved biological activity, enhanced pharmacokinetics, or innovative scaffolds. [4–8]

Although the concept of active machine learning has been conceived more than 30 years ago [9–11] and has been applied in select drug discovery campaigns for more than 15 years [12], its broader deployment to drug discovery pipelines across academia and industry has been surprisingly slow. [2] Possible reasons for these delays were attributed to two separate challenges. Firstly, infrastructural incompatibilities with rigid experimental high-throughput technologies popular among drug discovery facilities often made the deployment of such adaptive methods unfeasible. [13,14] Furthermore, the

introduction of active machine learning coincided with a general disenchantment of computational methods for drug discovery caused by unmet expectations that hampered trust and investments [15,16]. With improving flexibility of automation technology and a re-discovered enthusiasm for the deployment of artificial intelligence across drug discovery pipelines, active learning is expected to see a surge of applications [3,17,18]. To ensure that the setup of these pipelines is done with the greatest chance of success, it is important to consider the conclusions from previously published applications of active machine learning to drug discovery and other chemistry optimizations. This review provides a summary of the most important practical considerations for discovery teams interested in installing active learning by aggregating findings from previous applications of active learning to drug discovery campaigns.

**Implementing the right workflows**

A crucial component of an active learning workflow is the selection and training of a suitable machine learning model. Virtually all flavors of currently popular algorithms have been implemented in active learning pipelines (Table 1), including random forest models [4,5,19], Gaussian processes [20,21], support-vector machines [12,22,23], and (deep) neural networks [24,25]. This is encouraging since it suggests the applicability of a wide range of machine learning models to active learning, thereby enabling research teams to augment their machine learning model of choice. Whether any model architecture is particularly suited to actively learn a specific drug discovery challenge, similarly to large-scale benchmarking studies in classical QSAR, will be the subject of future studies. Even more importantly, a surge of recent papers have advocated for the utility of alternative learning approaches such as meta-learning [26], transfer-learning [27], multi-task learning [28], few-shot learning [29], as well as generative models [30]. Their integration within active learning pipelines might lead to strong synergies with even further improved accuracy or data economy.

While machine learning model selection does not appear to critically impact learning efficiency, the implemented data selection function strongly influences learning trajectories, model improvement, and quality of retrieved molecular material [2,6,21,31]. Commonly, active learning selection functions are designed to select the data least understood by the model in an effort to add new knowledge to the training data. This notion can be formalized through mathematical formulations of predictive uncertainty. Commonly, predictive uncertainty is quantified from the predictive variance across an ensemble of models [5,6,23,31] or through a distance measure to a decision boundary. [22,32] An opportunity exists to derive more complex formulations of utility for an experiment, for example by directly considering the predictive architecture of the model. [2] Further studies will be necessary to understand in which scenarios such more complex functions could provide any advantages over simpler models [5,21,33].

Particularly powerful active learning campaigns in drug discovery will account for the multi-objective character of its molecular optimizations while enabling the integration of orthogonal biochemical data and information. This could, for example, include notions of chemical or experimental tractability [4,7] or include chemistry-focused measures of anticipated novelty such as scaffold-diversity or physical simulations of potential binding modes [5,6,8]. Through such tailored integrations, active learning can specifically support drug discovery pipelines beyond simple data-driven optimizations. Furthermore, benchmarking implementations for success as defined by medicinal chemistry guidelines, such as

improved retrieval of novel active chemotypes [4–6] or robustness against enriching false assay positives [31] can aid selecting implementations that appear most promising for drug discovery applications [2].

**When to start and when to stop learning**

One of the first question a drug discovery team seeking to implement active learning will face is when to start an active learning campaign to best support a project. In essence, the question revolves around whether active learning should be implemented right at the beginning of a project and be used in the acquisition of the very first data, or whether there is value in harnessing historic data to augment the model. A general trend seems to be that multiple of the prospective active learning studies rely on providing an initial training data set (Table 1) [4,5,7] while the retrospective studies investigate active learning behavior when starting without prior data [6,31]. To date, no systematic comparison has been made and it is not entirely clear whether different initialization strategies impact the transferability of drawn conclusions across studies. Active learning performance is dependent on model quality [21,34], which potentially questions the utility of "cold-starting" active learning. However, analysis of learning trajectories has shown that even in the first iterations, active learning will select experiments in a more balanced manner compared to other strategies such as random or greedy sampling [35]. Furthermore, it is clear that different initial training datasets impact active learning behavior [23,25,36] while it seems distinct "cold-started" active learning campaigns are overall consistent and tractable [6,31,35]. While more evaluations are necessary to better understand the implications of different starting points for active learning campaigns, such decisions are expected to be mostly driven by the availability of resources and quality of prior data. While there certainly might be a reporting bias, the good news is that all active learning implementations have shown a significant benefit for projects following adaptive experimental design early or later in the study [2].

A similar practical question revolves around when to stop the active learning campaign. Tightly connected to this question is our ability to anticipate the performance of our current model and understanding the benefits of acquiring additional data. It has been shown that external data can be practically employed to either track the performance of the currently learned machine learning model [22] or be used to estimate expected hit rates in future screening iterations [37]. While these formalizations provide useful tools to aid adjusting expectations, the external data requirements might be difficult to meet and therefore of limited practical relevance. Instead of relying on external data, it has been shown that machine learning and simulated data can be harnessed to anticipate the performance of the current machine learning model [36,38]. Similarly, tracking changes in model architecture or predictive confidence on unlabeled data can assist in making decisions on when additional active learning iterations appear ineffective [4,5,36]. Researchers have also investigated opportunities to assess the expected increase in model performance from additional rounds of active learning: multiple studies have consistently shown that active learning curves typically indicate an exponential decay of error [6,22,23,31,34] and researchers have used analytical modelling and statistics to estimate the benefits of adding additional training data [31,35]. Through such efforts, transparency of expected performance and necessary resources can be created.

In practice, rather than relying on sophisticated estimations of learning rates, most active learning campaigns will be terminated either because of depleted resources or because a certain goal, such as the identification of a desirable molecular solution (Figure 2), has been achieved [1,7,38]. Additionally, multiple recent active learning platforms have highlighted the potential to not exclusively focus on

explorative learning but adaptively switch selection strategies or balance selection according to multiple explorative and exploitative objectives [4,5,24]. Thereby, rather than halting learning and changing selection towards hit identification, such hybrid platforms can continuously learn and adapt their behavior according to prospective hit rates [2,39]. It has been argued that such platforms will benefit from the multi-objective selection criteria and focus learning on the most relevant regions of chemical space [5,24]. Therefore, such platforms can be expected to see the largest amount of traction in deployment and will shape automated molecular optimizations in the future.

**The infrastructure bottlenecks**

One of the key discrepancies between most active learning conceptualizations and the practical reality of biochemical testing is the sequential character of active learning contrasting the parallelization of experimentation [2,6]. Virtually all *in vitro* experiments have been sufficiently miniaturized to enable the rapid testing of multiple hypotheses simultaneously. In fact, most experimental protocols and equipment are designed to capture multiple samples rather than testing one-by-one, making sequential single experiments unfeasible. Additionally, with increasing model complexities and associated training costs, some active learning workflows have found it unfeasible to re-train the model architecture for every new data point included [25,40]. Taken together, many workflows will have experimental or computational necessities to select multiple experiments with a single machine learning model before the model can be updated. Unfortunately, naively adding the top candidates has been shown to lead to redundancies in the selected experiments [5] and thereby significantly decrease active learning performance [6]. Recent active learning research has therefore focused on formalizations to improve batch selection of experiments with various proposed strategies. For example, researchers have actively forced diversity by restricting the sampling of the active learning function into poorly understood subsets [4,8] or by decreasing the density of the investigated experimental space through subsampling [7]. Instead of restricting the design space, another promising strategy is to iteratively regularize the active learning selection, either by grouping of experiments [38], through assessing the similarity of experimental parameters [32], or by consulting the model architecture to estimate perceived differences of potential experiments [5]. Alternatively, multiple distinct selection functions can be defined that steer the selection of independent experiments and thereby avoid redundancy [8,24]. While all these different approaches have been shown to reduce redundancy and boost active learning performance when batch selection is necessary, it is not entirely clear what the advantages of these distinct strategies are. Further evaluations will be necessary to directly compare such strategies and identify guidelines to pick the best batch selection method for a specific project.

Even if careful batch selection is implemented, a major hurdle in the deployment of active learning workflows remains the rigidness of high-throughput platforms commonly implemented across pharmaceutical and biotechnological companies. Such platforms are carefully designed and optimized to enable rapid screening of pre-defined compound libraries, for example through pre-plating of collections and installation of robotic automation that is implemented to quantitatively increase the throughput of such campaigns. However, these setups prevent adaptive cherry-picking of individual compounds. Thereby, the most valuable experiments according to active learning objectives is not individually accessible without significant overhead. This has prevented the deployment of adaptive machine intelligence and instead shifted the focus of software development on high-throughput data analysis [41]. However, the increasing awareness of distinct needs across projects has made improving

flexibility and adaptability of high-throughput testing platforms a key objective of compound management and experimental high-throughput technologies [13,17,42,43].

Accordingly, innovative experimental platforms that implement the required flexibility have enabled some of the most significant active learning studies published (Table 1). Naik *et al*. developed a pipeline where individual compounds were distributed with help of a liquid handling station and effects were analyzed via automated microscopy [38]. Granda *et al*. used a flow-based platform with 27 pumps to drive an eight chamber reaction platform with in-line NMR analytics to automatically discover novel chemistry [44]. Desai *et al*. integrated in-line synthesis, purification, and biological testing in a microfluidic platform that is driven by random forest-based active learning to discover novel Abl kinase inhibitors [4]. While the fully-integrated character of these platforms is a truly impressive technological advancement, their suitability for active learning campaigns is exclusively determined by their ability to adapt experimental design after every performed experiment. Multiple systems have been implemented that follow less automated approaches and streamline experimental design involving more manual labor [5,6,19,32]. These will enable the deployment of active learning in a wide range of settings without advanced laboratory hardware and broaden the scope of machine learning-driven optimization. Nevertheless, given the ultimate goal of automation to reduce manual intervention for improved throughput and reproducibility, active learning-driven automation will likely play a key role in future drug discovery efforts.

**Adjusting expectations**

It is clear that active learning is only one of many experimental design technologies that have been applied successfully in the drug discovery context. For example, diversity selections [45] and iterative screenings [43] are popular approaches to compound management and high-throughput screening. Some small-scale comparisons have shown that active learning might enable a more fine-tuned approach that adjusts to prior data and can be programmed to more rapidly home in on promising solutions [6,23]. Another popular method in the chemical science is factorial design, providing experimental guidelines to explore the impact of different parameters systematically. However, many relevant challenges in the chemical sciences have increasingly large parameter spaces that cannot be effectively enumerated. If active learning reaches a similar acceptance and becomes easily accessible through innovative software solutions [46,47], it might provide a competitive option for optimizing chemistry on complex and high-dimensional response surfaces. A small number of studies have compared performance of active learning campaigns to human optimizations and have found that active learning not only outcompetes the queried experts but also performed optimizations in a more systematic and explorative manner [19,32]. While the number of queried human experts in these studies is yet too small to draw definite conclusions, the consistently observed benefit of active learning in independent studies is promising. Further studies will be necessary to fully delineate the advantages of active learning and other experimental design approaches in different use cases, but it appears as if active machine learning is certainly ready to perform automated optimization campaigns with at least competitive outcomes to other approaches.

Large scale retrospective analysis have shown that active learning can identify highly accurate machine learning models using between half and down to one order of magnitude less data compared to classical machine learning and data subsampling approaches (Figure 3) [22,31,34,36,40]. Although the reasons for this higher efficiency are not yet completely understood, it seems that reduced redundancy and bias

as well as acquiring more meaningful data to span decision boundaries are major factors in this improved performance [12,22,31]. Thereby, without taking into account potential technological overhead for physically implementing active learning campaigns, it appears that screening costs can be reduced by at least 50% and up to 90%. Hypothetically, the costs could even be further decreased if concepts like cost-sensitive learning are considered [2]. Importantly, through this increased efficiency, sampling of larger parameter spaces becomes feasible and thereby potentially provides improved solutions [40]. With these benefits in mind, it is important to point out that performance will continue to vary widely from project to project [22,31]. Future method development efforts will need to focus on defining applicability domains for active learning workflows and provide transparency about expected performance [48]. This will be particularly challenging in the context of active learning, since this technology is particularly attractive for cases with limited data. Without data, however, *a priori* feasibility assessments will be challenging [49]. Admittingly, this challenge applies to any predictive technology applied to a novel use case, and active learning might eventually prevail as an adaptive approach, enabling rapid adjustment of experimental protocols according to improving understanding of the underlying design challenges [39,50].

**Conclusions**

With increasing deployment of flexible laboratory automation [4,17,44] and given the recent re-discovered enthusiasm for machine learning applications in drug discovery and development, active machine learning will become a key technology to guide molecular optimizations [2,18]. The set of previously published active learning applications serve as guideposts to inform future pipeline deployment [4–6,38]. These publications have outlined clear benefits of adaptive machine learning, both in terms of model improvement [4,38] as well as in the quality of retrieved molecular material [5,6]. A decreasing amount of necessary prior data (Table 1) and an increasing inclusion of orthogonal data and computations, including physical simulations and pharmacokinetic predictions, are trends likely to gain further traction in automated molecular optimizations [5,7,8]. Key methodological developments will have to delineate benefits of complex (batch) selection approaches [2] as well as defining the applicability domain and the anticipated benefits of active learning for a broad range of different applications [35,48]. Available implementations of active learning will simplify deployment without the need for re-implementing and re-validating code for individual projects [46,47]. Furthermore, integration of innovative learning approaches such as multitask-learning or generative models can potentially generate strong synergies in the future. Overall, active learning and related algorithmic tools are expected to qualitatively improve the reproducibility, throughput, and robustness of future drug discovery pipelines and provide an important tool in the search for innovative molecular solutions.

**Acknowledgements**

Daniel Reker is supported by the Swiss National Science Foundation (grants P2EZP3_168827 and P300P2_177833), the MIT-IBM Watson AI Lab, and the MIT SenseTime alliance.

**Table 1**: Examples of prospective active learning studies applied in drug discovery settings for the identification of small molecular probes with desired biological activity. Iterations corresponds to the number of compound selected through the active learning.

| Target(s) | Prior data | Iterations | Model | Infrastructure | Ref. |
|---|---|---|---|---|---|
| GPCRs | 20,000 | 10 | QBag | Manual | Fujiwara 2008 [6] |
| GPCRs | 215,967 | 29 | Bayesian | Manual | Besnard 2012 [7] |
| Abl kinase | 36 | 90 | Random forest | Flow | Desai 2013 [4] |
| CXCR4 | 287 | 90 | Random forest | Manual | Reker 2016 [5] |
| Protein localizaton | 96 | 1670 | Structure learning | Liquid handling | Naik 2016 [38] |

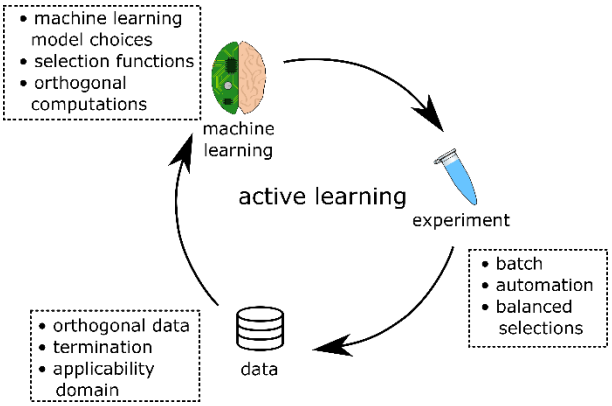


**Figure 1:** Active learning concept and key practical considerations in the setup of a novel active learning workflow.

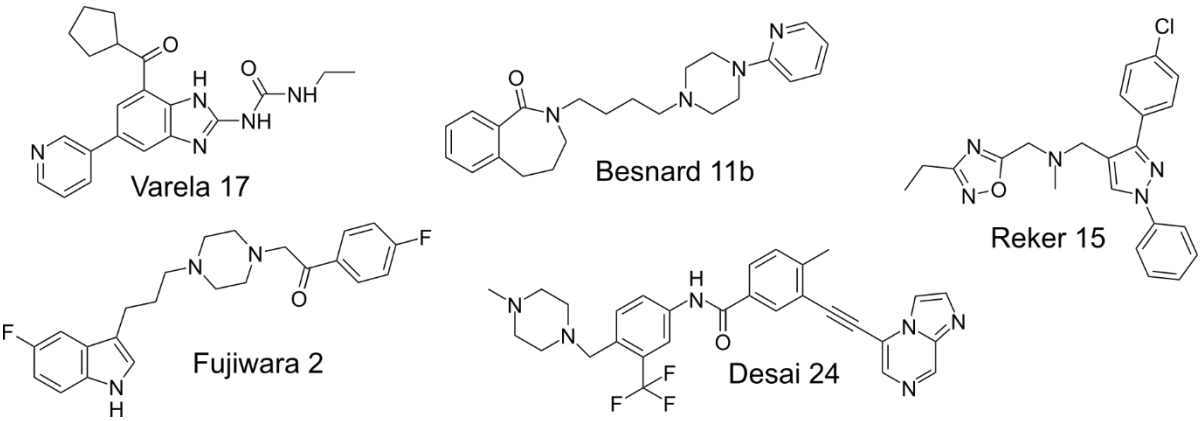


**Figure 2:** Molecular structures identified through active learning against a range of different therapeutic targets [4–8].

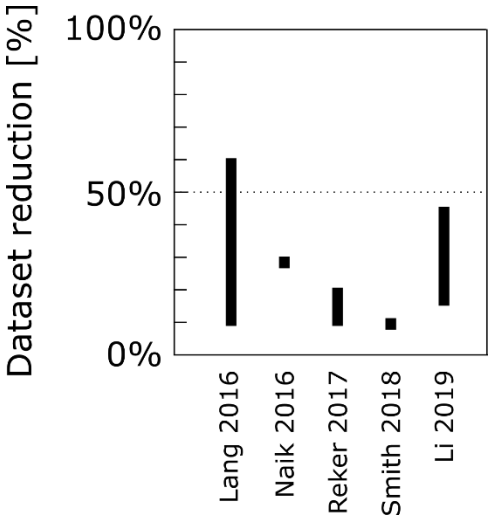


**Figure 3:** Expected benefit of active learning in terms of dataset reduction. Results are aggregated from a broad range of different active learning implementations and applications [22,31,36,38,40].

**References**

[1] Settles B. Active Learning. vol. 6. Morgan & Claypool Publishers; 2012. https://doi.org/10.2200/S00429ED1V01Y201207AIM018.

[2] Reker D, Schneider G. Active-learning strategies in computer-assisted drug discovery. Drug Discov Today 2015;20:458–65. https://doi.org/http://dx.doi.org/10.1016/j.drudis.2014.12.004.

[3] Murphy RF. An active role for machine learning in drug development. Nat Chem Biol 2011;7:327–30.

[4] Desai B, Dixon K, Farrant E, Feng Q, Gibson KR, van Hoorn WP, et al. Rapid discovery of a novel series of Abl kinase inhibitors by application of an integrated microfluidic synthesis and screening platform. J Med Chem 2013;56:3033–47.

[5] 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 2016;7:3919–27. https://doi.org/10.1039/C5SC04272K.

[6] Fujiwara Y, Yamashita Y, Osoda T, Asogawa M, Fukushima C, Asao M, et al. Virtual screening system for finding structurally diverse hits by active learning. J Chem Inf Mod 2008;48:930–40.

[7] Besnard J, Ruda GF, Setola V, Abecassis K, Rodriguiz RM, Huang X-P, et al. Automated design of ligands to polypharmacological profiles. Nature 2012;492:215–20.

[8] Varela R, Walters WP, Goldman BB, Jain AN. Iterative refinement of a binding pocket model: active computational steering of lead optimization. J Med Chem 2012;55:8926–42.

[9] MacKay DJC. Information-Based Objective Functions for Active Data Selection. Neural Comput 1992;4:590–604. https://doi.org/10.1162/neco.1992.4.4.590.

[10] Cohn DA, Ghahramani Z, Jordan MI. Active learning with statistical models. J Artif Intell Res

1996;4:129–45. https://doi.org/10.1613/jair.295.

[11] Zhang BT, Veenker G. Neural networks that teach themselves through genetic discovery of novel examples. 1991 IEEE Int. Jt. Conf. Neural Networks, 1992, p. 690–5. https://doi.org/10.1109/ijcnn.1991.170480.

[12] Warmuth MK, Liao J, Rätsch G, Mathieson M, Putta S, Lemmen C. Active learning with support vector machines in the drug discovery process. J Chem Inf Comput Sci 2003;43:667–73. https://doi.org/10.1021/ci025620t.

[13] Janzen WP, Popa-Burke IG. Advances in improving the quality and flexibility of compound management. J Biomol Screen 2009;14:444–51. https://doi.org/10.1177/1087057109335262.

[14] Bleicher KH, Böhm H-J, Müller K, Alanine AI. Hit and lead generation: beyond high-throughput screening. Nat Rev Drug Discov 2003;2:369–78. https://doi.org/10.1038/nrd1086.

[15] Dearden JC, Cronin MTD, Kaiser KLE. How not to develop a quantitative structure–activity or structure–property relationship (QSAR/QSPR). SAR QSAR Environ Res 2009;20:241–66. https://doi.org/10.1080/10629360902949567.

[16] Brown N, Ertl P, Lewis R, Luksch T, Reker D, Schneider N. Artificial Intelligence in Chemistry and Drug Design - A Perspective. J Comput Aided Mol Des 2020;in press.

[17] Schneider G. Automating drug discovery. Nat Rev Drug Discov 2018;17:97–113. https://doi.org/10.1038/nrd.2017.232.

[18] Eisenstein M. Active machine learning helps drug hunters tackle biology. Nat Biotechnol 2020;38:512–4.

[19] Reker D, Bernardes G, Rodrigues T. Evolving and Nano Data Enabled Machine Intelligence for Chemical Reaction Optimization. Chemrxiv 2018. https://doi.org/10.26434/chemrxiv.7291205.v1.

[20] De Grave K, Ramon J, De Raedt L. Active learning for primary drug screening. Benelearn 08, Annu. Belgian-Dutch Mach. Learn. Conf., vol. 2008, 2008, p. 55–6.

[21] Ahmadi M, Vogt M, Iyer P, Bajorath J, Fröhlich H. Predicting potent compounds via model-based global optimization. J Chem Inf Mod 2013;53:553–9.

[22] Lang T, Flachsenberg F, Von Luxburg U, Rarey M. Feasibility of Active Machine Learning for Multiclass Compound Classification. J Chem Inf Mod 2016;56:12–20. https://doi.org/10.1021/acs.jcim.5b00332.

[23] Fusani L, Cabrera AC. Active learning strategies with COMBINE analysis: new tricks for an old dog. J Comput-Aided Mol Des 2019;33:287–94. https://doi.org/10.1007/s10822-018-0181-3.

[24] Häse F, Roch LM, Kreisbeck C, Aspuru-Guzik A. Phoenics: A Bayesian Optimizer for Chemistry. ACS Cent Sci 2018;4:1134–45. https://doi.org/10.1021/acscentsci.8b00307.

[25] Zhang Y, Lee A. Bayesian semi-supervised learning for uncertainty-calibrated prediction of molecular properties and active learning. Chem Sci 2019. https://doi.org/10.1039/C9SC00616H.

[26] Finn C, Abbeel P, Levine S. Model-agnostic meta-learning for fast adaptation of deep networks. 34th Int. Conf. Mach. Learn. ICML 2017, vol. 3, 2017, p. 1856–68.

[27] Pan SJ, Yang Q. A survey on transfer learning. IEEE Trans Knowl Data Eng 2010;22:1345–59.

https://doi.org/10.1109/TKDE.2009.191.

[28] Unterthiner T, Mayr A, Klambauer G, Steijaert M, Wegner J, Ceulemans H, et al. Deep learning as an opportunity in virtual screening. Adv Neural Inf Process Syst 2014;27.

[29] Snell J, Swersky K, Zemel R. Prototypical networks for few-shot learning. Adv. Neural Inf. Process. Syst., vol. 2017–Decem, 2017, p. 4078–88.

[30] Segler MHS, Kogej T, Tyrchan C, Waller MP. Generating Focused Molecule Libraries for Drug Discovery with Recurrent Neural Networks. ACS Cent Sci 2018;4:120–31. https://doi.org/10.1021/acscentsci.7b00512.

[31] Reker D, Schneider P, Schneider G, Brown J. Active learning for computational chemogenomics. Future Med Chem 2017;9:381–402. https://doi.org/10.4155/fmc-2016-0197.

[32] Duros V, Grizou J, Xuan W, Hosni Z, Long D-L, Miras HN, et al. Human versus Robots in the Discovery and Crystallization of Gigantic Polyoxometalates. Angew Chem Int Ed 2017;56:10815–20. https://doi.org/10.1002/anie.201705721.

[33] De Grave K, Ramon J, De Raedt L. Active learning for high-throughput screening. Discov. Sci. Conf., Springer; 2008, p. 185–96.

[34] Rakers C, Reker D, Brown JB. Small Random Forest Models for Effective Chemogenomic Active Learning. J Comput Aided Chem 2017;8:124–42.

[35] Reker D, Brown JB. Selection of Informative Examples in Chemogenomic Datasets. Methods Mol. Biol., 2018. https://doi.org/10.1007/978-1-4939-8639-2_13.

[36] Li B, Rangarajan S. Designing compact training sets for data-driven molecular property prediction. ArXiv Prepr ArXiv190610273 2019.

[37] Buendia R, Kogej T, Engkvist O, Carlsson L, Linusson H, Johansson U, et al. Accurate Hit Estimation for Iterative Screening Using Venn–ABERS Predictors. J Chem Inf Mod 2019;59:1230–7. https://doi.org/10.1021/acs.jcim.8b00724.

[38] Naik AW, Kangas JD, Sullivan DP, Murphy RF. Active machine learning-driven experimentation to determine compound effects on protein patterns. ELife 2016;5. https://doi.org/10.7554/eLife.10047.

[39] Donmez P, Carbonell JG, Bennett PN. Dual strategy active learning. Mach. Learn. ECML 2007, Springer; 2007, p. 116–27.

[40] Smith JS, Nebgen B, Lubbers N, Isayev O, Roitberg AE. Less is more: Sampling chemical space with active learning. J Chem Phys 2018. https://doi.org/10.1063/1.5023802.

[41] Malo N, Hanley JA, Cerquozzi S, Pelletier J, Nadon R. Statistical practice in high-throughput screening data analysis. Nat Biotechnol 2006;24:167–75. https://doi.org/10.1038/nbt1186.

[42] Paricharak S, Ijzerman AP, Bender A, Nigsch F. Analysis of Iterative Screening with Stepwise Compound Selection Based on Novartis In-house HTS Data. ACS Chem Biol 2016;11:1255–64. https://doi.org/10.1021/acschembio.6b00029.

[43] Mayr LM, Bojanic D. Novel trends in high-throughput screening. Curr Opin Pharm 2009;9:580–8. https://doi.org/10.1016/j.coph.2009.08.004.

[44] Granda JM, Donina L, Dragone V, Long DL, Cronin L. Controlling an organic synthesis robot with machine learning to search for new reactivity. Nature 2018;559:377–81. https://doi.org/10.1038/s41586-018-0307-8.

[45] Meinl T, Ostermann C, Berthold MR. Maximum-score diversity selection for early drug discovery. J Chem Inf Mod 2011;51:237–47.

[46] Green DVS, Pickett S, Luscombe C, Senger S, Marcus D, Meslamani J, et al. BRADSHAW: a system for automated molecular design. J Comput-Aided Mol Des 2019:1–19. https://doi.org/10.1007/s10822-019-00234-8.

[47] Danka T, Horvath P. modAL: A modular active learning framework for Python 2018.

[48] Rakers C, Najnin RA, Polash AH, Takeda S, Brown JB. Chemogenomic Active Learning's Domain of Applicability on Small, Sparse qHTS Matrices: A Study Using Cytochrome P450 and Nuclear Hormone Receptor Families. ChemMedChem 2018;13:511–21. https://doi.org/10.1002/cmdc.201700677.

[49] Sahigara F, Mansouri K, Ballabio D, Mauri A, Consonni V, Todeschini R. Comparison of Different Approaches to Define the Applicability Domain of QSAR Models. Molecules 2012;17:4791–810. https://doi.org/10.3390/molecules17054791.

[50] Baram Y, El-Yaniv R, Luz K. Online choice of active learning algorithms. JMLR 2004;5:255–91.