



Artificial Intelligence in Drug Discovery and Health Care Systems: Machine Learning Applications in Modern Pharmacy and Patient Care

¹Dr. K. Rama Krishna Baba

Faculty, Dr. B. R. Ambedkar Department of Legal Studies
Vijayawada, Andhra Pradesh, India
ORCID: 0009-0006-9123-4719

²Tanya Chopra

Assistant Professor, Fashion Design
PCTE Group of Institutes
Ludhiana, Punjab, India
ORCID: 0009-0004-0685-5103

³Dr. J. James

Assistant Professor, Faculty of Law
Dr. M. G. R. Educational and Research Institute, Deemed to be University
Chennai, Tamil Nadu 600095, India

⁴Neehar Anand

First-Year MBBS Student
Dayanand Medical College & Hospital
Ludhiana, Punjab, India

Abstract

Artificial intelligence (AI) is transforming the value chain in pharmaceuticals from molecular design to delivery of care but most available literature is based on narratives, focused on discovery and care separately, and gives a single accuracy score without including calibration, utility, and fairness evidence as required by the contemporary standards and regulators. This paper introduces AI-PHARM, a consistent machine-learning framework for evaluation with five common layers of data, representation, learning, evaluation and governance applied consistently to both ends of the value chain with the same methodology using publicly available recent datasets. In the drug-discovery track, six model families were compared on all six in vitro Absorption-Distribution-Metabolism-Excretion (ADME) endpoints of the publicly available Biogen dataset (3,521 compounds). The heterogeneous stacked ensemble model was found to be the most accurate on all endpoints (average $R^2 = 0.533$), significantly outperforming the best individual model on three endpoints, and applicability domain analysis showed increasing prediction error for new-to-chemistry compounds. For the pharmacists' screening of dysglycemia based on patient-care, five different classifiers were tested using 229,781 unique adult patients' records in the CDC Diabetes Health Indicators Dataset; a well-calibrated gradient boosting model yields an AUC ROC of 0.814 and near-perfect calibration (expected calibration error = 0.003) and positive net clinical benefit according to decision curve analysis, and a subgroup audit identifies differences in error rates due to population differences in terms of age and income. An example of a negative result—a finding of no value in applying the ensemble in the classification problem at hand—is given by the framework's fundamental idea: Increased model complexity needs to be supported empirically, and discrimination, calibration, clinical usefulness, and fairness need to be considered together. Publicly available datasets and an explicitly defined experiment design have been used for all experiments.

Keywords: artificial intelligence; drug discovery; ADME prediction; machine learning; pharmacy practice; clinical risk prediction; stacked ensemble; model calibration; algorithmic fairness

1. Introduction

Despite the many improvements and advances in the drug development process, pharmaceuticals are still amongst the most costly and unreliable products that humanity invents, and the pharmacy is where medicines and patients finally come together. There is a very long chain of choices between the two—from the choice of the target to work on, to which molecule to synthesize, which compound to develop, which patient to treat, and how to monitor that patient—all of which used to depend on expert judgment and empirical analysis. However, AI is now introduced in all stages of that chain. Structural prediction tools like AlphaFold 3 make predictions about the protein-ligand complex structure with an incredible accuracy (Abramson et al., 2024); explainable deep learning approaches discovered new classes of antibiotics (Wong et al., 2024; Liu et al., 2023); an inhibitor of TNIK developed with the use of generative AI is currently undergoing the trials for idiopathic pulmonary fibrosis (Ren et al., 2025); and large language models can provide expert-level answers to medical licensing questions (Singhal et al., 2025). These very technologies have brought about a change in patient management by introducing alert systems, foundation models based on electronic health records and imaging data, and pharmacist decision support tools (Adams et al., 2022; Jiang et al., 2023; Zhou et al., 2023; Chalasani et al., 2023).

1.1 Background

Traditional drug discovery is a 10-year process that costs billions of dollars and fails to yield an approved medicine in the majority of cases due to unfavorable absorption, distribution, metabolism and excretion (ADME) properties or unexpected toxicity (Zhang et al., 2025). In vitro ADME tests such as microsomal stability, permeability and efflux, solubility and plasma protein binding are therefore conducted on thousands of compounds in each programme and shape the design–make–test–analyze cycle (Fang et al., 2023). AI models based on historical ADME assay data could thus help triage virtual compounds before synthesizing them and have thus made ADME prediction one of the most practical uses of AI in industry (Swanson et al., 2024; Heid et al., 2024). Regarding patients, community and hospital pharmacists are the most available health-care providers for many groups of people and increasingly carry out screening, medication review and chronic disease management. Models that are accurately calibrated and fair can therefore allow extending those services to large groups of people inexpensively (Chalasani et al., 2023; Rajpurkar et al., 2022).

1.2 Need of Research

This study is driven by three gaps. The first one is that there is a lot of descriptive literature in the AI-in-pharmacy field where the reviews discuss the opportunities and hardly ever experiment with the methods using an independent, open-source dataset (Blanco-González et al., 2023; Askar et al., 2023). Secondly, the discovery-stage and care-stage applications have been mostly studied separately despite being exposed to similar methodological risks of dataset shift, optimistic validation, improper calibration, and opacity of behavior (Sahiner et al., 2023; Wornow et al., 2023). Thirdly, new guidelines like TRIPOD+AI demand the evidence on calibration, utility, and subpopulations' performance, while regulators released their first guidelines on how AI should be used in drug development (U.S. Food and Drug Administration [FDA], 2025; European Union, 2024). Therefore, a reproducible evaluation-focused framework applicable to both ends of the pharmaceutical process addresses a methodological gap.

1.3 Motivation

1. Molecules discovered by AI have been successful in achieving Phase I with a rate of 80-90%, much higher than the historic average of the industry; however, the success rate for Phase II is similar to those of traditional programmes (Jayatunga et al., 2024); better pre-clinical property prediction is one of the ways of decreasing attrition.

2. ADME databases that come with prospective industrial testing, like Biogen Public dataset (Fang et al., 2023), enable transparent benchmarking of model families on real-world chemical data.
3. Pharmacists have an opportunity to offer population-level screening for type 2 diabetes and pre-diabetes, but they need tools that will provide risk predictions that are accurate, calibrated, and fair for different populations (Chen et al., 2023).
4. Doctoral level research in the field of pharmaceutical AI needs not only accuracy metrics but also the uncertainty statistics, applicability domain analysis, decision utility, and fairness audit.

1.4 Challenges

- Lack of data and data heterogeneity: assay data consists of only several thousands of noisy observations, and clinical data is imbalanced and confounded (Deng et al., 2023).
- Data shift: novel chemical series and novel population of patients differ from training data (Sahiner et al., 2023).
- Interpretability: explanations are as important for pharmacists, chemists and regulators as predictions themselves (Wong et al., 2024).
- Calibration and fairness: a model that is good at ranking patients might consistently overestimate or underestimate risks for some subpopulations (Chen et al., 2023).
- Evaluation and governance: benchmarking leakage and unrealistic predictions have been observed even for state-of-the-art models (Buttenschoen et al., 2024).

1.5 Influencing Factors

- Representations: fingerprints, physiochemical descriptors, graph representations, chemical language models, and 3D representations (Ross et al., 2022; Heid et al., 2024).
- Algorithm families: tree ensembles are very competitive when dealing with tabular data, while deep learning methods excel when using very big or unstructured data (Grinsztajn et al., 2022).
- Quality and quantity of data: curation of open-access resources, for example, The Therapeutics Data Commons (Huang et al., 2022) and the MIMIC-IV database (Johnson et al., 2023).
- Clinical perspective: model utility at the point of care is determined by the choice of threshold, prevalence, and workflow (Collins et al., 2024).
- Regulation & ethics: risk-based approach to AI development in light of the EU AI Act and FDA's draft guidance, WHO guidance on big multimodal models (World Health Organization [WHO], 2024).

1.6 Conventional Methodologies

The traditional paradigm of drug discovery is based on high throughput screening, QSAR based on linear regression, and medicinal chemistry heuristics. Traditional methods for assessing risks for patient care include manually calculated scores and logistic regressions with only a few predictors. They have a good interpretability level but lack certain capabilities: linear QSAR cannot model non-additive structure-property dependencies, while manually calculated scores cannot account for interaction between risk factors. ML extends traditional methods by introducing non-linear models and combinations of models at the expense of complexity and governance (Zhang et al., 2025; Wang et al., 2023).

1.7 Contributions

The contributions of this paper are four-fold. (i) The AI-PHARM framework, a dual track, five layer architecture based on the same ML approach centered around evaluation to pre-clinical ADME prediction and pharmacist-driven patient risk stratification. (ii) Benchmarking of six families of models across the

six outcomes of the Biogen publicly available ADME dataset (Fang et al., 2023) along with an investigation of when a stacked heterogeneous ensemble would help. (iii) Testing of five classifiers across 229,781 unique adult records of the CDC Diabetes Health Indicators dataset (Centers for Disease Control and Prevention [CDC], 2023) using calibration, decision curve, and subgroup-fairness analyses following TRIPOD+AI guidelines. (iv) Full protocol for the analysis is being made public to reproduce each number presented here from publicly available data.

2. Literature Review

Huang et al. (2022) developed the Therapeutics Data Commons, which is an integrated library of curated datasets, task-based splits and leaderboards for target discovery, ADMET, drug-target interaction, and outcome prediction of clinical trials. Standardisation of tasks and scaffold-based split strategy by the resource increased the reproducibility of comparisons among ML approaches for therapeutics but does not provide prospective leaderboard performance.

Ross et al. (2022) presented MoLFormer – transformer chemical language model trained on approximately 1.1 billion SMILES and found that the embeddings obtained encode molecular structure and transfer to downstream tasks of predicting properties. It became evident in the work that self-supervised pre-training is promising in chemistry and needs large compute budget.

Rajpurkar et al. (2022) surveyed AI progress in health and medicine and discussed the transition from retrospective proof-of-concept to prospective deployments with the focus on data access, label quality and workflow integration as the most significant bottlenecks. Adams et al. (2022) provided one of the few prospective evidences: timely clinician confirmation of alerts produced by TREWS sepsis early warning system decreased in-hospital mortality risk.

Fang et al. (2023) performed an analysis of the performance of machine learning models for six in vitro ADME endpoints over 20 months at Biogen and found that gradient-boosting and deep-learning models outperformed random forests and that frequent retraining performed better than hyperparameter optimization. Notably, they published 3,521 different commercial compounds tested on the same assays, the dataset used in Track A of this study.

Deng et al. (2023) performed a comprehensive study of important factors in molecular property prediction, training thousands of models and proving that representation-learning models frequently cannot outperform fixed fingerprints in small datasets and that activity cliffs and the size of the dataset determine the maximum possible accuracy. Heid et al. (2024) discussed Chemprop, a popular directed message-passing neural network software package, and Swanson et al. (2024) developed ADMET-AI from Chemprop to screen large compound libraries against 41 Therapeutics Data Commons ADMET endpoints.

In another study, Liu et al. (2023) employed a message-passing neural network to sift through 7,500 molecules to identify abaucin, an antibiotic with a narrow spectrum of activity against *Acinetobacter baumannii*. Similarly, Wong et al. (2024) adopted an advanced machine learning strategy to discover a class of antibiotics that were active against methicillin-resistant *Staphylococcus aureus* in mice using an explainable graph-based substructure rationale; both studies exemplify how ML may be used to accelerate the hit identification process.

Lin et al. (2023) found out that the ESM-2 protein language model could be used for atom-level structure prediction (ESMFold), whereas Watson et al. (2023) proposed RFdiffusion, which is capable of designing new protein binders from scratch. In addition, Abramson et al. (2024) adapted the AlphaFold algorithm to predict the structure of proteins, nucleic acids, and small-molecule ligands jointly. It is important, however, to note that many deep learning-based docking strategies produce physically invalid protein-ligand complex structures and fail to generalize to novel protein sequences.

Jayatunga et al. (2024) studied the clinical pipelines of AI-native biotech companies and found success rates of Phase I ranging between 80-90%, higher than the historical average of about 40-65%, while Phase II success rate (around 40%) was similar to the industry average. In another paper, Ren et al. (2025) detailed how the TNIK inhibitor INS018_055 (rentosertib) was discovered using generative AI for target identification and molecule design and provided Phase I safety information. Zhang et al. (2025) reviewed AI from drug development pipeline aspects from target identification to clinical trials and post-market surveillance.

With regard to healthcare applications, Jiang et al. (2023) used NYUTron pre-trained on unstructured clinical notes and demonstrated that one health system language model could predict readmission, mortality, comorbidity, and insurance denial. Zhou et al. (2023) developed RETFound which is self-supervised foundation model for retinal image and its generalization ability for disease screening. Chen et al. (2024) introduced UNI, a general computational pathology foundation model. The study of Singhal et al. (2025) demonstrated that Med-PaLM 2 has expertise to provide answers to questions but Wornow et al. (2023) pointed out that foundation models are evaluated on narrow clinical tasks.

In the context of pharmacy practice, Chalasani et al. (2023) conducted a review on AI in drug information services, prescription verification, adverse drug reactions detection and medication adherence, and Salas et al. (2022) conducted a systematic review on AI in pharmacovigilance, noting that most of the available literature is proof of concept and does not undergo external validation. Chen et al. (2023) identified potential sources of bias in medical AI algorithms and possible solutions to this problem, and Collins et al. (2024) developed TRIPOD+AI, which mandates the reporting of calibration, clinical utility and fairness of prediction models using machine learning.

Table 1. Summary of Literature Review

S. No.	Author(s) & Year	Focus	Methodology / Data	Key Limitation
1	Huang et al. (2022)	Benchmark ecosystem for therapeutics ML	Curated datasets, scaffold splits, leaderboards	Retrospective only; no prospective test
2	Ross et al. (2022)	Chemical language model (MoLFormer)	Transformer pre-trained on ~1.1 B SMILES	Heavy compute; limited gain on small sets
3	Adams et al. (2022)	Sepsis early warning (TREWS)	Prospective multi-site cohort	Observational; single health system
4	Fang et al. (2023)	Prospective ADME prediction	120 internal datasets; 3,521-compound public set	Single company's assays and chemistry
5	Deng et al. (2023)	Drivers of molecular property prediction	Large-scale systematic model comparison	Mostly public benchmarks; few ADME assays
6	Liu et al. (2023)	Antibiotic discovery (abaucin)	Message-passing neural network + wet lab	Narrow-spectrum target organism
7	Jiang et al. (2023)	Clinical language model (NYUTron)	Health-system-scale clinical notes	Single institution; data not public
8	Chalasani et al. (2023)	AI in pharmacy practice	Narrative literature review	No quantitative evaluation
9	Wong et al. (2024)	Explainable antibiotic discovery	Graph neural networks + substructure rationales	Validation largely pre-clinical
10	Buttenschoen et al. (2024)	Validity of AI docking	PoseBusters physical-validity test suite	Docking only; small-molecule focus

11	Jayatunga et al. (2024)	Clinical success of AI drugs	Pipeline analysis of AI-native biotechs	Small sample; early-phase dominated
12	Collins et al. (2024)	Reporting standard (TRIPOD+AI)	Delphi consensus guideline	Guidance, not empirical evidence
13	Ren et al. (2025)	Generative-AI drug (TNIK inhibitor)	Target + molecule generation, Phase I	Single programme; company-reported
14	Singhal et al. (2025)	Medical LLM (Med-PaLM 2)	Benchmarks + physician evaluation	Question answering, not patient outcomes

3. Problem Statement

Although there have been advances, there is an evaluation issue for AI in drug development and healthcare that is shared. ADME models developed in the discovery phase are generally evaluated using a single measure of accuracy for one split, without uncertainty measures or an assessment of how performance deteriorates when faced with new chemical molecules. Risk models used in the care phase are evaluated purely on discrimination (AUROC), despite pharmacists acting based on probability predictions and threshold values, and despite performance varying by sex, age, and socioeconomic group. For both applications, there is no evidence on (a) which model family to use, (b) whether additional complexity, like ensembling, adds value, and (c) whether the model is safe and ethical to use. To that end, this study seeks to answer three research questions:

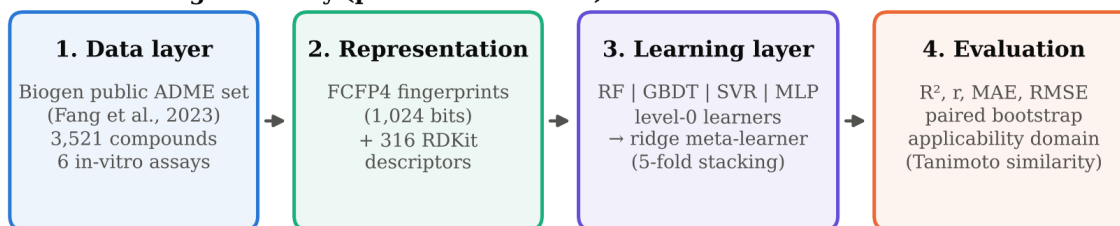
- RQ1: Which families of linear, kernel, tree-ensemble, neural, and stacked ensemble models perform best in predicting six in-vitro ADME endpoints from recently released public industrial data?
- RQ2: How effective are routine measures of health in identifying adults who have pre-diabetes or diabetes for pharmacist screening?

RQ3: Would these models still be accurate for chemically novel compounds and equitable for all socio-demographic sub-groups?

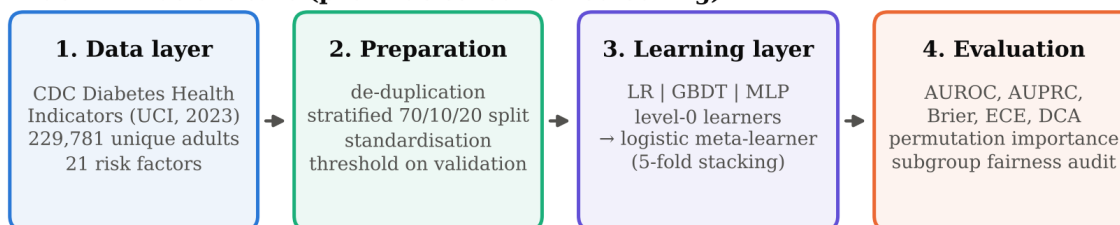
4. Proposed Work: The AI-PHARM Framework

We introduce AI-PHARM (AI for PHarmaceutical research and Medicines-use), an integrated evaluation-oriented framework in which the same machine learning approach is used for two opposite ends of the pharmaceutical value chain: preclinical property prediction in drug discovery (Track A) and risk profiling in patient care (Track B). The framework consists of five layers—data, representation, learning, evaluation, and governance—which span both tracks, as presented in Figure 1. The rationale is that the choice of models is less important than the evaluation process: each model is evaluated based on its accuracy and uncertainty, calibration, and applicability or fairness and utility, according to the most recent guidance (Collins et al., 2024).

Track A - Drug discovery (pre-clinical ADME)



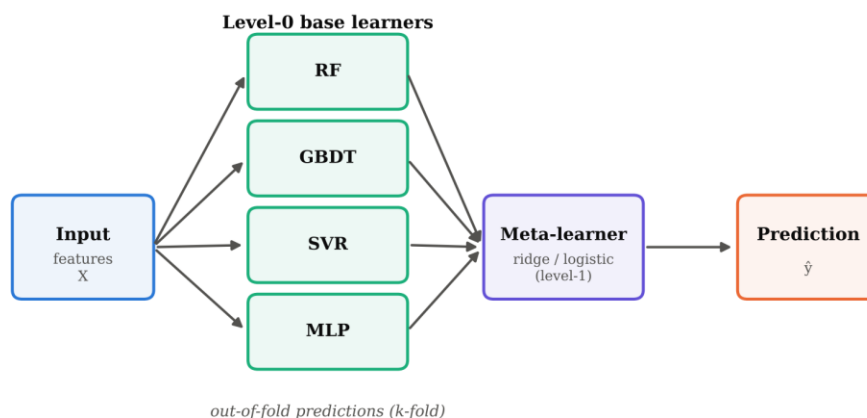
Track B - Patient care (pharmacist-led risk screening)



5. Governance layer: TRIPOD+AI reporting | explainability | fairness & calibration monitoring
 drift surveillance | FDA (2025) draft guidance and EU AI Act (2024) alignment

Figure 1. The AI-PHARM framework: two tracks (drug discovery and patient care) sharing five layers — data, representation, learning, evaluation and governance.

Architecture is illustrated in Figure 1. Data layer receives open recent datasets for each track. Representation layer transforms raw data into model-ready features, namely molecular fingerprints and descriptors for Track A, standardized clinical measures for Track B. Learning layer trains a group of heterogeneous base models that get combined with cross-validated stacked meta-learner (Figure 2). Evaluation layer utilizes track-specific metrics with bootstrap uncertainty. Governance layer wraps the entire process in reporting, explainability, fairness, and drift monitoring techniques consistent with FDA (2025) guidance and EU AI Act (European Union, 2024).



Base learners are trained with internal cross-validation; their out-of-fold predictions train the meta-learner, which combines them into the final output.

Figure 2. Stacked-generalisation architecture. Heterogeneous level-0 learners produce out-of-fold predictions that train a regularised level-1 meta-learner.

4.1 Datasets

Track A (drug discovery). We employ the Biogen publicly available ADME dataset from Fang et al. (2023), which includes 3,521 purchased compounds screened via six standard in-vitro assays according to

the authors' internal procedures: HLM intrinsic clearance, RLM intrinsic clearance, MDR1-MDCK efflux ratio (permeability/efflux proxy), aqueous solubility at pH 6.8, human and rat plasma-protein binding (hPPB, rPPB). All endpoints are modeled using the log-transformed values used in the dataset. The authors' released splits into training and test sets are adopted for all endpoints, ensuring no test compounds appear during training. The numbers of samples per endpoint are provided in Table 2.

Track B (patient care). We will use the CDC Diabetes Health Indicators dataset (CDC, 2023; UCI Machine Learning Repository), which comes from the 2015 Behavioral Risk Factor Surveillance System. In total, there are 253,680 survey records for adults containing 21 risk indicators and a three-level glycemic state label (no diabetes, pre-diabetes, diabetes). After filtering out 23,899 duplicate records, there are 229,781 distinct respondents left. In accordance with the pharmacists' case-finding problem, where the task was to identify patients who need further testing for diabetes, we constructed a binary target, indicating either pre-diabetes or diabetes (17.1% prevalence). There are 21 predictors in total including medical predictors (hypertension, hypercholesterolemia, BMI, stroke/coronary disease), behavioral (smoking, physical activities, diet, alcohol abuse), self-assessed general/physical/mental well-being, and socio-demographic predictors (gender, age group, education, income level).

4.2 Representation

In Track A, the molecules are encoded using the feature sets provided along with the dataset: the 1,024-bit functional-class fingerprint (FCFP4) feature set extended by 316 RDKit physicochemical descriptors, which constitutes 1,340 features. Fixed fingerprints along with descriptors are still a robust representation choice for small ADME datasets, where they can even compete with machine learning-derived representations (Deng et al., 2023). Missing and non-finite descriptors are imputed with their medians, and the descriptors themselves are standardised prior to usage in scale-sensitive learners. In Track B, the 21 indicators are used as they are, whereas continuous variables are standardised in case of linear and neural learners.

4.3 Learning Layer

Each task involves training a heterogeneous panel of base learners and a stacked ensemble (Figure 2). For regression (Track A), the panel includes ridge regression, random forest (RF), histogram gradient-boosted decision tree (GBDT), support vector machine with an RBF kernel (SVR), and multilayer perceptron (MLP, two hidden layers, target standardization and strong weight decay). For classification (Track B) – penalized logistic regression (LR), RF, GBDT, and MLP. The model of interest is stacked generalisation ensemble: the base learners represent level-0, out-of-fold predictions of base learners are generated via internal cross-validation, and a regularized meta-learner (ridge regression for regression, logistic regression for classification) combines predictions of base learners at level-1. Stacking is selected due to partial independence of mistakes made by base learners, which means that convex combination could help reducing variance without model selection. All hyperparameters are set a priori based on default values and light pilot tuning on training data only; no test data were used for modelling decisions.

In the case of Track B, in which there are five candidate models (the four base classifiers and the stacked ensemble) that can be chosen from, the operational model will be selected on the basis of the held out validation dataset by using the Brier score, in which both discrimination and calibration get considered; the threshold for the choice is determined at the point where Youden's J is maximised.

4.4 Evaluation Layer

For Track A, we provide the coefficient of determination (R^2), Pearson's r , Spearman's ρ , mean absolute error (MAE) and root-mean-square error (RMSE) calculated on the held-out test set of each endpoint. In order to measure uncertainty and perform model comparisons, we used a paired bootstrap (1,000 resamples of the test set) providing the mean and 95% confidence interval (CI) of the R^2 difference

between the stacked ensemble and each baseline. In order to investigate generalisability to new chemical space (applicability domain), we calculate, for each test molecule, the maximum Tanimoto similarity of its fingerprint to the training set, and analyse the relationship between the error and the similarity (Figure 5).

For Track B, we provide discrimination (AUROC, AUPRC), calibration (Brier score, expected calibration error [ECE] and reliability plots), and confusion matrix analysis (sensitivity, specificity, positive and negative predictive values, F1) at the chosen threshold. The CIs for AUROC are obtained using 500-sample bootstrap, and the proposed model is compared with each baseline using a paired bootstrap of the AUROC difference. Since discrimination alone is insufficient for demonstrating clinical utility, we use decision curve analysis (see Figure 8). Interpretability is evaluated using permutation importance (see Figure 9), while subgroup fairness audit (see Table 6) is done considering prevalence, AUROC, true-positive rate and false-positive rate for different sex, age bracket, income and education subgroups.

4.5 Governance Layer

Governance is the layer that makes responsible deployment work. The report is done using the TRIPOD+AI checklist (Collins et al., 2024). Explanation is given through permutation importance and applicability-domain analysis for Track A. Fairness and calibration are tracked through the subgroup audit and reliability diagrams. Since the medical and chemical data tend to drift over time (Sahiner et al., 2023), the framework includes revalidation and recalibration according to the findings of Fang et al. (2023), and consistent with the FDA (2025) risk-based credibility-assessment methodology and the high-risk system standards of the EU AI Act (European Union, 2024). All experiments rely on fixed random seeds and public datasets to be reproducible.

4.6 Experimental Setup

All model implementations are done using Python 3 and scikit-learn. The experiments were conducted in a 2-core Linux system. The random seeds are fixed to 42. In track A, the provided split per endpoint is used without changes. In track B, the de-duped data are split using stratified sampling into 70% training, 10% validation and 20% testing. No extra data or pre-training has been used. The full procedure is deterministic based on the provided inputs.

5. Results and Discussion

5.1 Track A: ADME Property Prediction

Table 2 shows the endpoints and the test set performance of the models. Amongst all the six endpoints, the stacked ensemble showed the highest R^2 values, lowest RMSE and lowest MAE, thus demonstrating that the stacking of heterogenous learners works well for ADME prediction (see Figure 3). The average R^2 value for the stacked ensemble amongst all the six endpoints was $R^2 = 0.533$, while the best performing family (GBDT) showed an R^2 value of 0.492, followed by SVR (0.476) and random forest (0.443). The relative difficulties of endpoints followed the expected trend, with plasma-protein binding and MDR1-MDCK efflux being the easiest (hPPB, $R^2 = 0.64$; MDR1-MDCK, $R^2 = 0.63$; rPPB, $R^2 = 0.56$), while aqueous solubility being the hardest to predict ($R^2 = 0.38$).



Figure 3. Track A test-set R^2 for every model and endpoint. The stacked ensemble is highest on all six endpoints.

Table 2. Track A — endpoints and test-set performance (R^2 , with RMSE in parentheses; best model per endpoint in bold)

Endpoint	n train / test	Ridge	RF	GBDT	SVR	MLP	Stacked (proposed)
HLM (log CL int)	2469 / 618	0.341 (0.503)	0.371 (0.491)	0.419 (0.472)	0.434 (0.466)	0.416 (0.473)	0.476 (0.448)
RLM (log CL int)	2443 / 611	0.403	0.427	0.472	0.453	0.400	0.515
MDR1-MDCK ER	2113 / 529	0.541	0.498	0.577	0.556	0.572	0.628
Solubility (pH 6.8)	1738 / 435	0.327	0.312	0.343	0.333	0.327	0.384
hPPB (% unbound)	1446 / 362	0.591	0.561	0.604	0.543	0.615	0.636
rPPB (% unbound)	708 / 177	0.469	0.460	0.537	0.428	0.528	0.559
Mean	—	0.445	0.443	0.492	0.458	0.476	0.533

The order of base learners is in line with that reported by Deng et al. (2023) and Fang et al. (2023), where the gradient boosting and kernel/neural learners were better than random forests, while the baseline ridge regression model performed well only for easier endpoints. There was no base learner dominating all scenarios, where SVR was superior to HLM, GBDT to RLM/MDR1/rPPB and MLP to hPPB.

In order to confirm the reliability of the ensemble advantage, the paired-bootstrap R^2 difference between the stacked ensemble and the best single model was calculated for each endpoint (Table 3). The ensemble was significantly better than the best single learner for three out of six endpoints (HLM, RLM, MDR1-MDCK; 95% CI not including zero) while showing a positive but insignificant advantage for the rest (Solubility, hPPB, rPPB), where the gradient-boosting and neural models performed almost as good as the ensemble model. In case of poorer baseline models (ridge, random forest), the advantage of the ensemble

was significant for all endpoints. This means that stacking gives a robust "no-regret" choice as it will be always at least as good as the best single model and will usually be significantly better at the price of around 10 times longer training time (116 s compared to 2–17 s per endpoint)..

Table 3. Track A — paired-bootstrap R^2 gain of the stacked ensemble over the best single model per endpoint (1,000 resamples)

Endpoint	Best single model	ΔR^2 (stacked – best single)	95% CI	Significant
HLM	SVR	+0.041	[0.009, 0.068]	Yes
RLM	GBDT	+0.043	[0.020, 0.067]	Yes
MDR1-MDCK ER	GBDT	+0.050	[0.028, 0.072]	Yes
Solubility	GBDT	+0.040	[−0.001, 0.082]	No (borderline)
hPPB	MLP	+0.022	[−0.014, 0.057]	No
rPPB	GBDT	+0.023	[−0.019, 0.068]	No

Figure 4 gives the plots of predicted versus measured parity plots for the stacked ensemble across all six endpoints; prediction is in line with the regression to the mean for the extreme values. The applicability domain analysis is shown in Figure 5: the mean absolute error decreases monotonically with increasing maximum Tanimoto similarity of a test molecule to the training set, ranging from 0.50 log units for the most novel molecules (similarity < 0.2) to 0.33 log units for the closest analogues (similarity $\geq$ 0.5). This illustrates an important limitation for use—predictions for novel chemical scaffolds are significantly less reliable—and justifies the recommendation to report a per-molecule confidence flag according to training set similarity, analogous to the data drift problem pointed out by Sahiner et al. (2023) and validation failures reported by Buttenschoen et al. (2024).

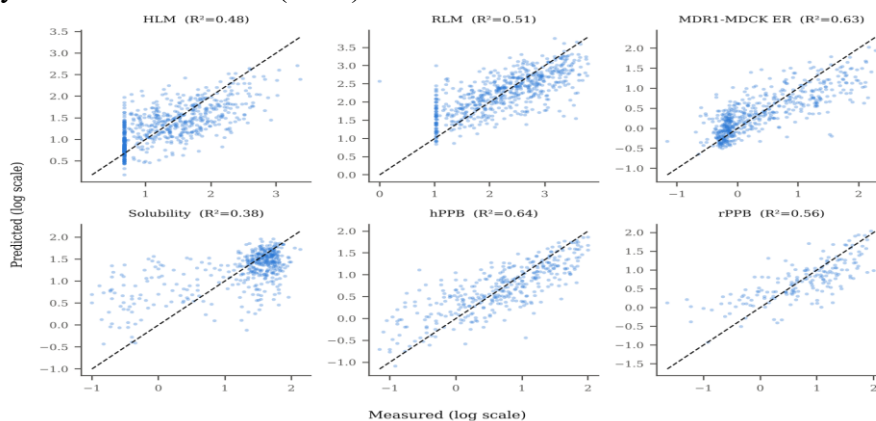


Figure 4. Track A predicted-versus-measured parity plots for the stacked ensemble; the dashed line is the identity.

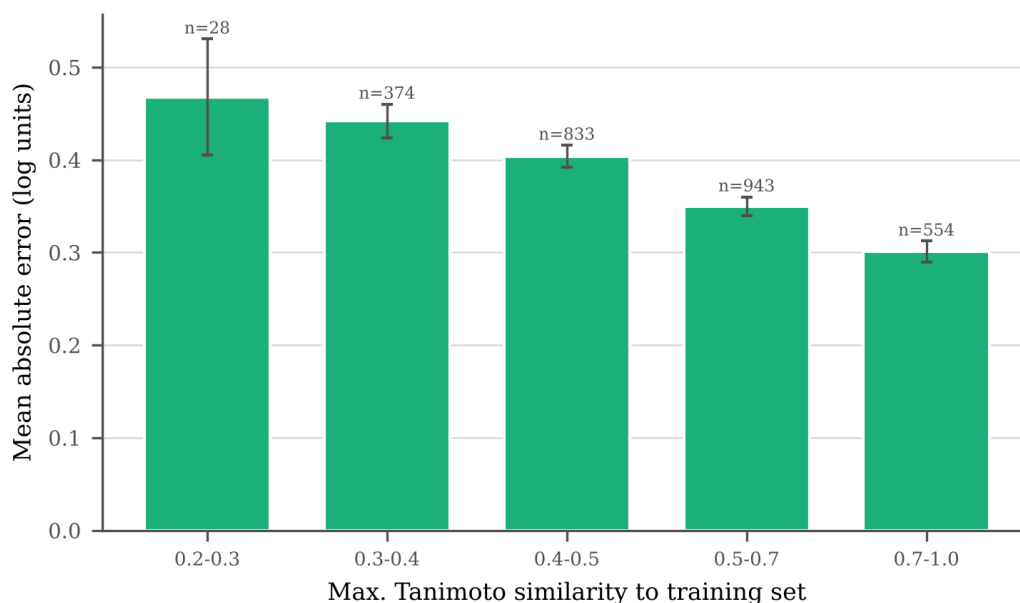


Figure 5. Track A applicability domain: mean absolute error of the stacked ensemble versus maximum Tanimoto similarity of the test compound to the training set (error bars: standard error).

5.2 Track B: Pharmacist-Led Dysglycaemia Screening

After deduplication, there were 229,781 unique adults (17.3% had pre-diabetes/diabetes: 4,629 pre-diabetes and 35,097 diabetes), which were divided into training (70%, $n=160,846$), validation (10%, $n=22,978$) and test sets (20%, $n=45,957$). The test set performance of the five models is provided in Table 4. All five models had comparable discrimination ($\text{AUROC} = 0.805\text{--}0.814$), but varied significantly in calibration. The GBDT model had the best validation Brier score, thus it was chosen as the final model; it showed $\text{AUROC} = 0.814$ (95%CI 0.810-0.819), $\text{AUPRC} = 0.472$ (prevalence 0.173), Brier = 0.115 and expected calibration error = 0.003.

Table 4. Track B — test-set performance ($n = 45,957$; selected model in bold)

Model	AUROC (95% CI)	AUPRC	Brier	ECE
Logistic regression	0.805 (0.800–0.809)	0.441	0.118	0.013
Random forest	0.811 (0.807–0.816)	0.468	0.115	0.012
GBDT (selected)	0.814 (0.810–0.819)	0.472	0.115	0.003
MLP	0.810 (0.806–0.815)	0.461	0.116	0.005
Stacked ensemble	0.814 (0.810–0.819)	0.471	0.116	0.029

It should be pointed out that, unlike Track A, the stacked ensemble does not perform better than the best individual model for this classification task; the stacked ensemble's AUROC is no higher than that of GBDT (+0.000 AUROC; paired-bootstrap 95% CI: -0.001 to $+0.001$; $p = 0.36$), and its calibration is worse (ECE 0.029 compared with 0.003). This is an interesting negative result. In the case of a large, low-dimensional tabular dataset, a well-regularized gradient-boosting model is enough to use all the

information contained in the data, which means that there are no errors left for a meta-learner to correct—this is consistent with the effectiveness of tree ensembles on tabular data noted by Grinsztajn et al. (2022). The ROC curves are displayed in Figure 6 and reliability diagrams in Figure 7, whereby the GBDT curve follows the diagonal very closely for the entire range of probabilities, but the stacked model exhibits over-prediction of probability on the high-probability side. At the validation-operating point (0.149, optimizing Youden's J statistic), the GBDT algorithm provided a sensitivity of 0.808, specificity of 0.660, positive predictive value of 0.332 and negative predictive value of 0.943. As an opportunistically used tool in the pharmacy for identifying people with dysglycaemia, such values are quite reasonable, since it can identify about four out of every five people with the condition whilst rejecting 66% of the negatives, and a negative prediction is correct 94% of the time.

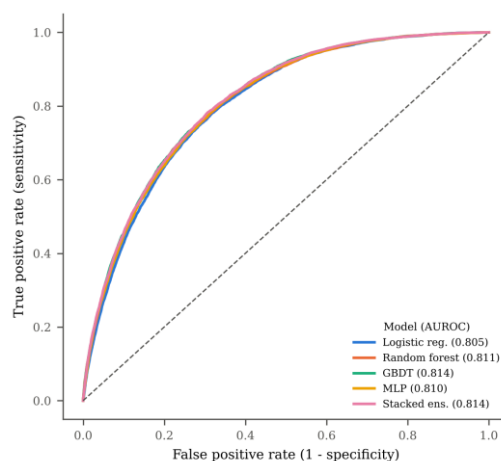


Figure 6. Track B receiver-operating-characteristic curves on the test set (AUROC in parentheses).

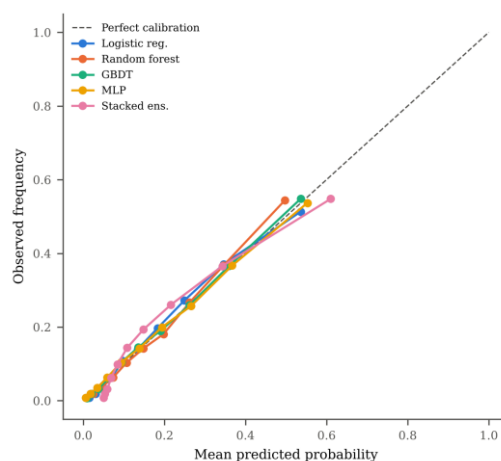


Figure 7. Track B reliability diagram; the selected GBDT model lies almost on the diagonal.

Usefulness in a clinical setting is evidenced via decision curve analysis (see Figure 8). Within the feasible screening range (threshold probabilities of 0.05-0.30), GBDT demonstrated the highest net benefit compared to all other models and was superior to "screen everyone" and "screen no one." At the threshold probability of 10%, the net benefit for GBDT was 0.113, while the net benefit for "screen everyone" was 0.081—roughly corresponding to an increase in the detection rate of almost three more actual cases out of every hundred patients screened but without an increase in the number of redundant confirmations.

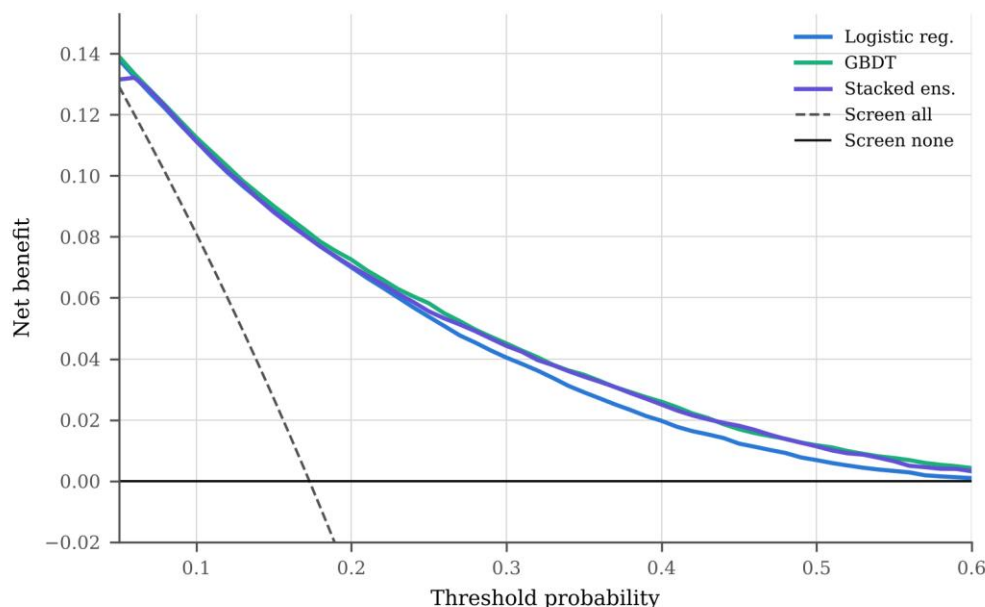


Figure 8. Track B decision-curve analysis. The selected GBDT model gives the highest net benefit across the screening range and dominates screen-all and screen-none.

Permutation importance for the selected model is depicted in Figure 9. The top predictors include self-rated overall health status, body mass index, age group, high blood pressure, and high blood cholesterol level. This list can be considered clinically sensible and consistent with the known risk factors for diabetes, which may allow using a concise questionnaire instead.

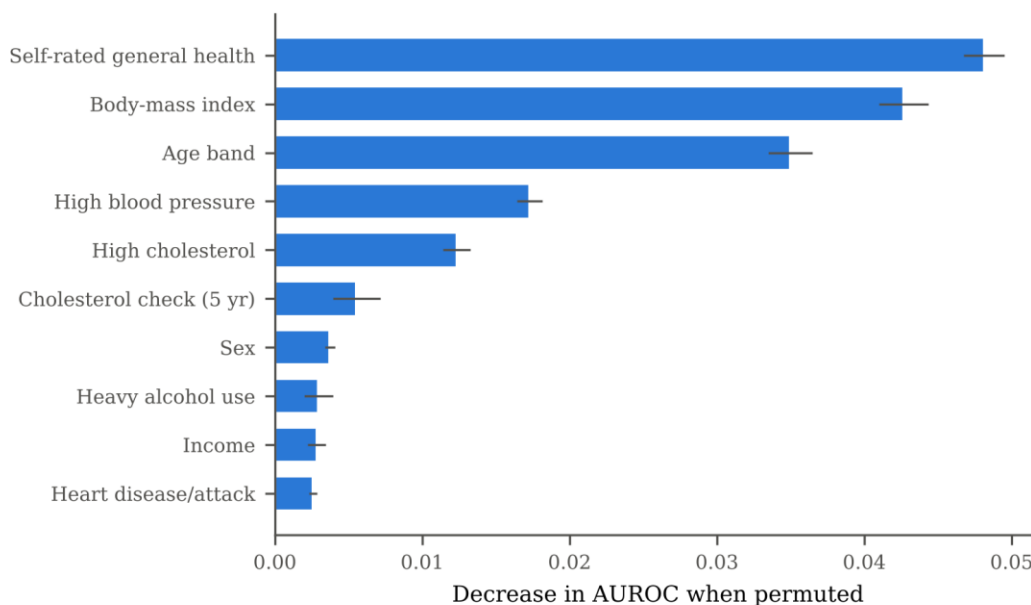


Figure 9. Track B permutation importance (top 10 predictors) for the selected GBDT model.

5.3 Fairness and Subgroup Analysis

The Subgroup Audit table (Table 5) shows the results of the selected GBDT model's audit by sex, age, income, and education levels. There was little discrimination, with AUROC within subgroups being between 0.762 and 0.823. Moreover, the mean predicted probability matched the prevalence rate in all the subgroups, meaning that calibration was maintained for each stratified population. Still, the false positive

rate was very uneven – ranging from 0.04 for the youngest and least prevalent subgroup to 0.54 for the oldest and least wealthy ones – since one global cut-off interacts with different prevalence rates among the subgroups. The true positive rate demonstrated the opposite trend. The example demonstrates the existing tension of fairness in modelling (Chen et al., 2023) – a well-calibrated model may have an uneven distribution of error rates among different groups.

Table 5. Track B — subgroup performance of the selected GBDT model at the global threshold

Attribute	Group	n	Prevalence	AUROC	TPR	FPR
Sex	Female	25,671	0.163	0.823	0.800	0.310
Sex	Male	20,286	0.186	0.802	0.816	0.378
Age	18–39 y	6,981	0.038	0.815	0.358	0.043
Age	40–59 y	16,407	0.143	0.821	0.726	0.251
Age	≥60 y	22,569	0.236	0.762	0.866	0.528
Income	< \$25k	7,367	0.276	0.784	0.907	0.537
Income	\$25k–<\$75k	16,046	0.193	0.800	0.837	0.406
Income	≥ \$75k	22,544	0.124	0.814	0.702	0.243
Education	≤ High school	14,620	0.224	0.796	0.868	0.447
Education	Some college	13,459	0.173	0.814	0.805	0.347
Education	College graduate	17,878	0.131	0.816	0.726	0.257

5.4 Discussion

There are two clear lessons to take away from reading the results side-by-side. The first is that the very same discipline – using a diverse set of base learners, stacking meta-learners, and performance measurement going well beyond just one number for accuracy – translates directly from molecules to patients, reinforcing the argument for a common framework across discovery and treatment. The second is that the usefulness of the suggested method depends on the task at hand, and needs to be proven, rather than presumed – while stacking yielded statistically significant improvements on the small, noisy, high-dimensional problems of ADME prediction, it showed no improvement on the large, low-dimensional classification problem, where a calibrated gradient boosting model had already proved optimal.

This highlights another reason to consider evaluation outside of discrimination. In Track A, the applicability domain analysis showed that while headline R^2 hides the substantial reliance on chemical novelty; in Track B, there were two competing models which differed by nearly ten times in terms of calibration error despite the identical AUROCs, and while the one which demonstrated the greatest discrimination power was not chosen to be deployed. When it comes to drugs, whose prediction is used for making the synthesis and patient referral decisions, calibration, clinical usefulness and subgroup behavior are not extra but rather essential characteristics of the model to be deployed.

Lastly, the research serves to confirm a general observation made by notable research findings in recent times, namely that AI is already creating real value in the pharma domain, from drug discovery (Ren et al., 2025) to new antibiotics (Wong et al., 2024), to increased early-phase clinical trial success (Jayatunga et al., 2024); however, there have been several examples in the same period of warning findings regarding benchmark validity (Buttenschoen et al., 2024) and physical reality (Wornow et al., 2023). An evaluation-focused approach such as the AI-PHARM is one solution to this conflict.

5.5 Limitations

Several shortcomings exist here. For Track A, the results are dependent on the particular company's assay conditions, along with a commercially skewed space of chemicals, hence, absolute accuracies will not necessarily carry over to other labs; in addition, we have used the published random splits and not scaffold splits that tend to overestimate the performance for novel chemicals; the applicability-domain analysis has partially taken care of this issue. We have used fixed fingerprints-plus-descriptors representations without testing graph neural networks and chemical language models (Heid et al., 2024; Ross et al., 2022) which could alter the rankings on larger datasets. For Track B, the data consist of self-reported surveys from 2015 where the label is generated from the survey itself as opposed to laboratory-verified glycaemic status, there is a mix of both pre-diabetes and diabetes in the positive class, and the cohort is of U.S. origin, hence, external and time validation are required prior to any deployment. The hyperparameters were fine-tuned only lightly and we performed no prospective analysis since Fang et al. (2023) and Adams et al. (2022) have shown that this constitutes the decisive factor.

6. Conclusion

In this paper, we presented AI-PHARM, a coherent and evaluation-focused machine learning framework bridging between two extremes of the pharmaceutical pipeline, namely pre-clinical ADME prediction in drug discovery and pharmacist-based risk screening among patients, and evaluated it on two recent publicly available datasets. In our evaluation using Biogen's public ADME dataset, a stacked heterogeneous ensemble delivered highest accuracies for all six in-vitro outcomes ($R^2 = 0.533$ on average) with statistically significant improvements over the best individual model in three endpoints; applicability domain analysis helped us quantify the deterioration of performance for novel chemical structures. In our experiments using the CDC Diabetes Health Indicators dataset, a gradient boosting model, properly calibrated, predicted pre-diabetes and diabetes with AUROC = 0.814, near-perfect calibration (ECE = 0.003), and obvious net health benefit, while subgroup audit revealed error rate differences driven by prevalence that need to be handled prior to deployment. The negative results which have been deliberately communicated—that there was no benefit from the ensemble in the classification task—serve to emphasize the crucial principle in the framework, that model complexity should be demonstrated through empirical evidence, and that discrimination, calibration, clinical usefulness, and fairness need to be considered in concert. In employing a single methodological approach to discovery and treatment, and publishing an entirely reproducible methodology based on publicly available data, this study provides an exemplary approach to AI in contemporary pharmacy.

7. Future Scope

Future work will take five directions. Firstly, the ADME track will be generalized from existing molecular representations to directed message passing graph neural networks and chemical language models pre-trained (Heid et al., 2024; Ross et al., 2022), and to scaffold split and time split evaluation using Therapeutics Data Commons ADMET benchmarks (Huang et al., 2022), in order to evaluate generalization to novel chemistry. Secondly, the patient care track will be validated in an external setting for laboratory-confirmed cohorts and more recent survey waves, and will be further extended to cover additional conditions relevant to community pharmacy practice, such as cardiovascular risk and drug harm. Thirdly, uncertainty quantification (conformal prediction) and per-prediction explanation will be integrated to give both the chemist and the pharmacist not only the confidence but also the reason behind each prediction. Fourthly, prospective, embedded workflow evaluation—by far the most convincing demonstration carried out by Fang et al. (2023) and Adams et al. (2022)—will evaluate actual decision impact, not just accuracy⁵. Finally, the governance layer will be implemented via drift detection, scheduling for calibration, and creation of a credibility assessment dossier in line with the FDA (2025)

draft guidance on explainability of artificial intelligence medical software as a product and the EU AI Act (European Union, 2024).

References

- 1 Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., Ronneberger, O., Willmore, L., Ballard, A. J., Bambrick, J., Bodenstein, S. W., Evans, D. A., Hung, C.-C., O'Neill, M., Reiman, D., Tunyasuvunakool, K., Wu, Z., Žemgulytė, A., Arvaniti, E., ... Jumper, J. M. (2024). Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, 630(8016), 493–500. <https://doi.org/10.1038/s41586-024-07487-w>
- 2 Adams, R., Henry, K. E., Sridharan, A., Soleimani, H., Zhan, A., Rawat, N., Johnson, L., Hager, D. N., Cosgrove, S. E., Markowski, A., Klein, E. Y., Chen, E. S., Saheed, M. O., Henley, M., Miranda, S., Houston, K., Linton, R. C., Ahluwalia, A. R., Wu, A. W., & Saria, S. (2022). Prospective, multi-site study of patient outcomes after implementation of the TREWS machine learning-based early warning system for sepsis. *Nature Medicine*, 28(7), 1455–1460. <https://doi.org/10.1038/s41591-022-01894-0>
- 3 Askr, H., Elgeldawi, E., Aboul Ella, H., Elshaier, Y. A. M. M., Gomaa, M. M., & Hassanien, A. E. (2023). Deep learning in drug discovery: An integrative review and future challenges. *Artificial Intelligence Review*, 56(7), 5975–6037. <https://doi.org/10.1007/s10462-022-10306-1>
- 4 Blanco-González, A., Cabezón, A., Seco-González, A., Conde-Torres, D., Antelo-Riveiro, P., Piñeiro, Á., & Garcia-Fandino, R. (2023). The role of AI in drug discovery: Challenges, opportunities, and strategies. *Pharmaceuticals*, 16(6), 891. <https://doi.org/10.3390/ph16060891>
- 5 Buttenschoen, M., Morris, G. M., & Deane, C. M. (2024). PoseBusters: AI-based docking methods fail to generate physically valid poses or generalise to novel sequences. *Chemical Science*, 15(9), 3130–3139. <https://doi.org/10.1039/D3SC04185A>
- 6 Centers for Disease Control and Prevention. (2023). CDC Diabetes Health Indicators [Data set]. UCI Machine Learning Repository. <https://doi.org/10.24432/C53919>
- 7 Chalasani, S. H., Syed, J., Ramesh, M., Patil, V., & Pramod Kumar, T. M. (2023). Artificial intelligence in the field of pharmacy practice: A literature review. *Exploratory Research in Clinical and Social Pharmacy*, 12, 100346. <https://doi.org/10.1016/j.rcsop.2023.100346>
- 8 Chen, R. J., Ding, T., Lu, M. Y., Williamson, D. F. K., Jaume, G., Song, A. H., Chen, B., Zhang, A., Shao, D., Shaban, M., Williams, M., Oldenburg, L., Weishaupt, L. L., Wang, J. J., Vaidya, A., Le, L. P., Gerber, G., Sahai, S., Williams, W., & Mahmood, F. (2024). Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30(3), 850–862. <https://doi.org/10.1038/s41591-024-02857-3>
- 9 Chen, R. J., Wang, J. J., Williamson, D. F. K., Chen, T. Y., Lipkova, J., Lu, M. Y., Sahai, S., & Mahmood, F. (2023). Algorithmic fairness in artificial intelligence for medicine and healthcare. *Nature Biomedical Engineering*, 7(6), 719–742. <https://doi.org/10.1038/s41551-023-01056-8>
- 10 Collins, G. S., Moons, K. G. M., Dhiman, P., Riley, R. D., Beam, A. L., Van Calster, B., Ghassemi, M., Liu, X., Reitsma, J. B., van Smeden, M., Boulesteix, A.-L., Camaradou, J. C., Celi, L. A., Denaxas, S., Denniston, A. K., Glocker, B., Golub, R. M., Harvey, H., Heinze, G., ... Logullo, P. (2024). TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*, 385, e078378. <https://doi.org/10.1136/bmj-2023-078378>
- 11 Deng, J., Yang, Z., Wang, H., Ojima, I., Samaras, D., & Wang, F. (2023). A systematic study of key elements underlying molecular property prediction. *Nature Communications*, 14, 6395. <https://doi.org/10.1038/s41467-023-41948-6>

- 12 European Union. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union, L 2024/1689. <http://data.europa.eu/eli/reg/2024/1689/oj>
- 13 Fang, C., Wang, Y., Grater, R., Kapadnis, S., Black, C., Trapa, P., & Sciabola, S. (2023). Prospective validation of machine learning algorithms for absorption, distribution, metabolism, and excretion prediction: An industrial perspective. *Journal of Chemical Information and Modeling*, 63(11), 3263–3274. <https://doi.org/10.1021/acs.jcim.3c00160>
- 14 Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on typical tabular data? In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022), Datasets and Benchmarks Track* (pp. 507–520). Curran Associates. <https://doi.org/10.48550/arXiv.2207.08815>
- 15 Heid, E., Greenman, K. P., Chung, Y., Li, S.-C., Graff, D. E., Vermeire, F. H., Wu, H., Green, W. H., & McGill, C. J. (2024). Chemprop: A machine learning package for chemical property prediction. *Journal of Chemical Information and Modeling*, 64(1), 9–17. <https://doi.org/10.1021/acs.jcim.3c01250>
- 16 Huang, K., Fu, T., Gao, W., Zhao, Y., Roohani, Y., Leskovec, J., Coley, C. W., Xiao, C., Sun, J., & Zitnik, M. (2022). Artificial intelligence foundation for therapeutic science. *Nature Chemical Biology*, 18(10), 1033–1036. <https://doi.org/10.1038/s41589-022-01131-2>
- 17 Jayatunga, M. K. P., Ayers, M., Bruens, L., Jayanth, D., & Meier, C. (2024). How successful are AI-discovered drugs in clinical trials? A first analysis and emerging lessons. *Drug Discovery Today*, 29(6), 104009. <https://doi.org/10.1016/j.drudis.2024.104009>
- 18 Jiang, L. Y., Liu, X. C., Nejatian, N. P., Nasir-Moin, M., Wang, D., Abidin, A., Eaton, K., Riina, H. A., Laufer, I., Punjabi, P., Miceli, M., Kim, N. C., Orillac, C., Schnurman, Z., Livia, C., Weiss, H., Kurland, D., Neifert, S., Dastagirzada, Y., ... Oermann, E. K. (2023). Health system-scale language models are all-purpose prediction engines. *Nature*, 619(7969), 357–362. <https://doi.org/10.1038/s41586-023-06160-y>
- 19 Johnson, A. E. W., Bulgarelli, L., Shen, L., Gayles, A., Shammout, A., Horng, S., Pollard, T. J., Hao, S., Moody, B., Gow, B., Lehman, L.-W. H., Celi, L. A., & Mark, R. G. (2023). MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10, 1. <https://doi.org/10.1038/s41597-022-01899-x>
- 20 Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., dos Santos Costa, A., Fazel-Zarandi, M., Sercu, T., Candido, S., & Rives, A. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637), 1123–1130. <https://doi.org/10.1126/science.ade2574>
- 21 Liu, G., Catacutan, D. B., Rathod, K., Swanson, K., Jin, W., Mohammed, J. C., Chiappino-Pepe, A., Syed, S. A., Fragis, M., Rachwalski, K., Magolan, J., Surette, M. G., Coombes, B. K., Jaakkola, T., Barzilay, R., Collins, J. J., & Stokes, J. M. (2023). Deep learning-guided discovery of an antibiotic targeting *Acinetobacter baumannii*. *Nature Chemical Biology*, 19(11), 1342–1350. <https://doi.org/10.1038/s41589-023-01349-8>
- 22 Rajpurkar, P., Chen, E., Banerjee, O., & Topol, E. J. (2022). AI in health and medicine. *Nature Medicine*, 28(1), 31–38. <https://doi.org/10.1038/s41591-021-01614-0>
- 23 Ren, F., Aliper, A., Chen, J., Zhao, H., Rao, S., Kuppe, C., Ozerov, I. V., Zhang, M., Witte, K., Kruse, C., Aladinskiy, V., Ivanenkov, Y., Polykovskiy, D., Fu, Y., Babin, E., Qiao, J., Liang, X., Mou, Z., Wang, H., ... Zhavoronkov, A. (2025). A small-molecule TNIK inhibitor targets fibrosis in preclinical and clinical models. *Nature Biotechnology*, 43(1), 63–75. <https://doi.org/10.1038/s41587-024-02143-0>

- 24 Ross, J., Belgodere, B., Chenthamarakshan, V., Padhi, I., Mroueh, Y., & Das, P. (2022). Large-scale chemical language representations capture molecular structure and properties. *Nature Machine Intelligence*, 4(12), 1256–1264. <https://doi.org/10.1038/s42256-022-00580-7>
- 25 Sahiner, B., Chen, W., Samala, R. K., & Petrick, N. (2023). Data drift in medical machine learning: Implications and potential remedies. *The British Journal of Radiology*, 96(1150), 20220878. <https://doi.org/10.1259/bjr.20220878>
- 26 Salas, M., Petracek, J., Yalamanchili, P., Aimer, O., Kasthuril, D., Dhingra, S., Junaid, T., & Bostic, T. (2022). The use of artificial intelligence in pharmacovigilance: A systematic review of the literature. *Pharmaceutical Medicine*, 36(5), 295–306. <https://doi.org/10.1007/s40290-022-00441-z>
- 27 Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Amin, M., Hou, L., Clark, K., Pfohl, S. R., Cole-Lewis, H., Neal, D., Rashid, Q. M., Schaekermann, M., Wang, A., Dash, D., Chen, J. H., Shah, N. H., Lachgar, S., Mansfield, P. A., ... Natarajan, V. (2025). Toward expert-level medical question answering with large language models. *Nature Medicine*, 31(3), 943–950. <https://doi.org/10.1038/s41591-024-03423-7>
- 28 Swanson, K., Walther, P., Leitz, J., Mukherjee, S., Wu, J. C., Shivnaraine, R. V., & Zou, J. (2024). ADMET-AI: A machine learning ADMET platform for evaluation of large-scale chemical libraries. *Bioinformatics*, 40(7), btae416. <https://doi.org/10.1093/bioinformatics/btae416>
- 29 U.S. Food and Drug Administration. (2025). Considerations for the use of artificial intelligence to support regulatory decision-making for drug and biological products (Draft guidance for industry). U.S. Department of Health and Human Services. <https://www.fda.gov/regulatory-information/search-fda-guidance-documents>
- 30 Wang, H., Fu, T., Du, Y., Gao, W., Huang, K., Liu, Z., Chandak, P., Liu, S., Van Katwyk, P., Deac, A., Anandkumar, A., Bergen, K., Gomes, C. P., Ho, S., Kohli, P., Lasenby, J., Leskovec, J., Liu, T.-Y., Manrai, A., ... Zitnik, M. (2023). Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972), 47–60. <https://doi.org/10.1038/s41586-023-06221-2>
- 31 Watson, J. L., Juergens, D., Bennett, N. R., Trippe, B. L., Yim, J., Eisenach, H. E., Ahern, W., Borst, A. J., Ragotte, R. J., Milles, L. F., Wicky, B. I. M., Hanikel, N., Pellock, S. J., Courbet, A., Sheffler, W., Wang, J., Venkatesh, P., Sappington, I., Vázquez Torres, S., ... Baker, D. (2023). De novo design of protein structure and function with RFdiffusion. *Nature*, 620(7976), 1089–1100. <https://doi.org/10.1038/s41586-023-06415-8>
- 32 Wong, F., Zheng, E. J., Valeri, J. A., Donghia, N. M., Anahtar, M. N., Omori, S., Li, A., Cubillos-Ruiz, A., Krishnan, A., Jin, W., Manson, A. L., Friedrichs, J., Helbig, R., Hajian, B., Fiejtek, D. K., Wagner, F. F., Soutter, H. H., Earl, A. M., Stokes, J. M., ... Collins, J. J. (2024). Discovery of a structural class of antibiotics with explainable deep learning. *Nature*, 626(7997), 177–185. <https://doi.org/10.1038/s41586-023-06887-8>
- 33 World Health Organization. (2024). Ethics and governance of artificial intelligence for health: Guidance on large multi-modal models. World Health Organization. <https://www.who.int/publications/i/item/9789240084759>
- 34 Wornow, M., Xu, Y., Thapa, R., Patel, B., Steinberg, E., Fleming, S., Pfeffer, M. A., Fries, J., & Shah, N. H. (2023). The shaky foundations of large language models and foundation models for electronic health records. *npj Digital Medicine*, 6, 135. <https://doi.org/10.1038/s41746-023-00879-8>
- 35 Zhang, K., Yang, X., Wang, Y., Yu, Y., Huang, N., Li, G., Li, X., Wu, J. C., & Yang, S. (2025). Artificial intelligence in drug development. *Nature Medicine*, 31(1), 45–59. <https://doi.org/10.1038/s41591-024-03434-4>
- 36 Zhou, Y., Chia, M. A., Wagner, S. K., Ayhan, M. S., Williamson, D. J., Struyven, R. R., Liu, T., Xu, M., Lozano, M. G., Woodward-Court, P., Kihara, Y., Allen, N., Gallacher, J. E. J., Littlejohns,



T., Aslam, T., Bishop, P., Black, G., Sergouniotis, P., ... Keane, P. A. (2023). A foundation model for generalizable disease detection from retinal images. *Nature*, 622(7981), 156–163. <https://doi.org/10.1038/s41586-023-06555-x>