



Digital twin approaches for interpretable side effect prediction in drug discovery

András Ecker¹, Gergely Szabó¹, János Szalma^{1,2}, Erzsébet Fichó¹, István Reguly^{1,2}, Attila Csikász-Nagy^{1,2,*}

¹ Cytocast Hungary, Budapest, Hungary

² Faculty of Information Technology and Bionics, Pázmány Péter Catholic University, Budapest, Hungary

Artificial intelligence increasingly supports preclinical drug development although accurate, early prediction of side effects remains limited. Leading approaches rely on expert-derived features, unavailable in early discovery, or on chemical similarity-based graph neural networks functioning as ‘black boxes’ and lacking actionable insight. We propose a paradigm shift: instead of directly mapping chemical features to side effects, derive biologically meaningful intermediate representations, such as predicted off-target proteins and their downstream effects simulated in a cellular digital twin. Simple models trained on these representations are interpretable and therefore actionable. Beyond safety assessment, this approach could inform rational polypharmacology strategies and guide the amendment of secondary pharmacology assays.

Keywords: Drug discovery; side effects; machine learning; feature representation; interpretability; actionable insight

Introduction

Drug development is an intricate and risky process, often taking over a decade and costing billions of US dollars. Once efficacy is confirmed, the main cause for terminating clinical trials or even withdrawing drugs from the market is unwanted side effects. To alleviate the burden, several *in silico* methods have been established to predict potential side effects of drug candidates. In the 2010s, classical machine learning (ML) methods, such as logistic regression, support vector machines (SVMs), decision trees, matrix factorization, knowledge graph- and network-based methods, dominated the field.^{(p1),(p2)} These models combine features in a way that is easy to grasp and trivially invertible in the mathematical sense; thus the way their predictions are made can be interpreted. The features resulting in the best-scoring models include drug phenotype, for example the anatomical therapeutic chemical (ATC) code, and drug chemical features, for example Tanimoto similarity of Morgan fingerprints.^{(p1),(p2)} Whereas maximizing performance is important in

the ML community, and the interpretability of these algorithms is a huge plus, the interpretation of these features is not really actionable by practitioners. Predicting side effects that resemble those of previously approved drugs that belong to the same ATC category according to the WHO, and backing it by chemical similarity, is just not pragmatic.

When the predictive performance of these models plateaued in the 2020s, the focus shifted toward neural networks,^(p3) which gained traction in other biology-related domains, such as predicting the 3D structure of proteins.^(p4) However, to achieve their superior performance, deep neural networks require an immense amount of training data, which, in the case of adverse drug reactions (ADRs), are not publicly available. Furthermore, although neural networks build their own feature representation and mitigate the problem of feature engineering, without a thorough analysis of this representation, their predictions are much more difficult to interpret. This has earned them a ‘black box’ designation and the distrust of non-experts (Figure 1c).

* Corresponding author. Csikász-Nagy, A. (a.csikasz-nagy@cytocast.com)

Thus, with the exception of a few notable examples,^{(p5),(p6),(p7)} early-stage drug development to date does not obtain actionable insights from available side-effect prediction tools. Indeed, in the past year, the FDA and EMA pharmaceutical agencies came out with regulatory acceptance plans of new approach methodologies (NAMs), including *in silico* methods for side-effect predictions. Although these initiatives signal institutional interest and underscore the direction in which the field is evolving, their implementation will inevitably be accompanied by regulatory oversight and validation requirements. To date, no software has received such approval. In this Perspective, we argue that advancing the field requires training simpler models that leverage better features.

Simpler models using better features

Good feature representation

We will introduce the concept of good features with 2D classification tasks. When points belonging to two classes are detached blobs (Figure 1ai: the red and black dots), they can be easily separated by a line using their *x* and *y* coordinates as features. This can be done by using the two features simultaneously with a linear classifier, or by a more advanced SVM that maximizes the margins from the closest points (Figure 1aii, top). Another possible solution would be a decision tree that builds tree-like structures using if-then rules considering only a single feature at a time (Figure 1aii, bottom). Actual examples, however, tend to be more complicated and not easily separated using only the original input features. We illustrate this by classifying concentric circles (Figure 1b, left), which cannot be solved in 2D by a SVM. A large decision tree would achieve better performance but would overfit and be hard to interpret because of the high number of if-then decisions. By contrast, by introducing a new feature as the sum of squares, resulting in a paraboloid in 3D would make the problem again easily solvable (Figure 1b, right: blue hyperplane fitted by a SVM). This type of feature engineering is however time consuming and requires human intuition (Figure 1cii), which has led to the rapid rise of employing neural networks that build their own feature representation while solving the given task (Figure 1ci). When these networks were small, it was feasible to understand the features learned.^(p8) However, because their parameters exceeded the billions, there is no straightforward way to interpret them. In computer vision applications, it is a common practice to verify whether the layers of the network behave similarly to brain areas of the visual stream processing the same information.^(p9) In other fields like systems biology, however, this is simply not possible.

As for coming up with good feature representations, one can still turn to basic pharmacology. Notably, drug candidate compounds are designed to target specific proteins and exert their effects on these on-targets. The landmark study of Campillos *et al.*^(p10) showed that on-targets can be predicted from side effects. Although they did it for drug repositioning, it inspired some of the early side-effect prediction models to use on-targets as features.^(p2) By contrast, although even on-target perturbation can lead to undesired effects, most ADRs arise from unintentional binding to other, off-target proteins. Secondary

pharmacology assays are designed to test a few hundred important proteins linked to known side effects^(p11); thus, if available, these serve as good features. Indeed, the seminal work of Ietswaart *et al.*^(p6) used a plethora of these measurements to build XGBoost models (ensembles of decision trees) predicting ADRs. Because laboratory measurements are only available in a few cases, Sawada *et al.*^(p7) used predicted binding affinities against the human proteome (Figure 1d). They then predicted side effects with a simple logistic regression from these, which achieved higher accuracies than the same model trained on the same labels with traditional features. After clustering the predicted binding affinity matrix, they found that ATC codes were overrepresented in some of their clusters.^(p7) Similar conclusions were reached by Galeano *et al.*,^(p5) who used matrix decomposition to build low dimensional representations of drugs and side effects. Some of their 10D drug feature vectors were also significantly correlated with ATC codes. In summary, correlation with ATC codes that represent expert knowledge about the drug can be seen as a good sign of well-derived features. Note, however, that ATC codes themselves are only determined in the late phase of drug development (usually after clinical trials); thus they cannot be used to predict potential ADRs in the early phases of the pipeline.

For even more features, one can turn again to systems biology, and acknowledge the fact that the target proteins partake in a network of protein-protein interactions (PPIs), form protein complexes and participate in signaling pathways. Although powerful and widely used,^(p1) simple graph-based approaches on PPI networks neglect the fact that protein abundances and binding sites are limited. To take this into account and study complex formation under constrained resources, we built an agent-based simulator that mimics the behavior of proteins within the cell^(p12): the CYTOCAST DIGITAL TWIN Cell™ (Figure 2a, top). Relevant data were sourced from several databases: protein abundances from PaxDB,^(p13) protein domains from Pfam,^(p14) PPIs from STRING,^(p15) domain-domain interactions (DDIs) from 3did,^(p16) reference complexes from ComplePortal^(p17) and the Reactome Knowledgebase.^(p18) The simulator also tracks PPIs of signaling pathways, parsed from the Reactome Knowledgebase.^(p18) Although the simulations are stochastic, performing dozens of runs with different random seeds and doing statistical analysis on the distributions of runs makes the results as reliable as *in vitro* experiments. Running a simulation takes 30 core hours on AMD EPYC CPUs. Furthermore, in parallel with Sawada *et al.*,^(p7) we developed our own binding affinity predictor (unpublished) to predict targets of drug compounds from a list of 7390 druggable proteins (Figure 2a, b). Using these predicted targets we were able to extend the above simulations by adding interactions between the drug compound and domains of the predicted targets, which perturb the complexome and the signaling pathways (Figure 2a, bottom). These significantly perturbed complexes and pathways are highly preprocessed bio-realistic features, which can also be used for side-effect prediction. To derive them for a new compound of interest, only the compound structure is required; thus they are readily applicable in the earliest stage of drug development.

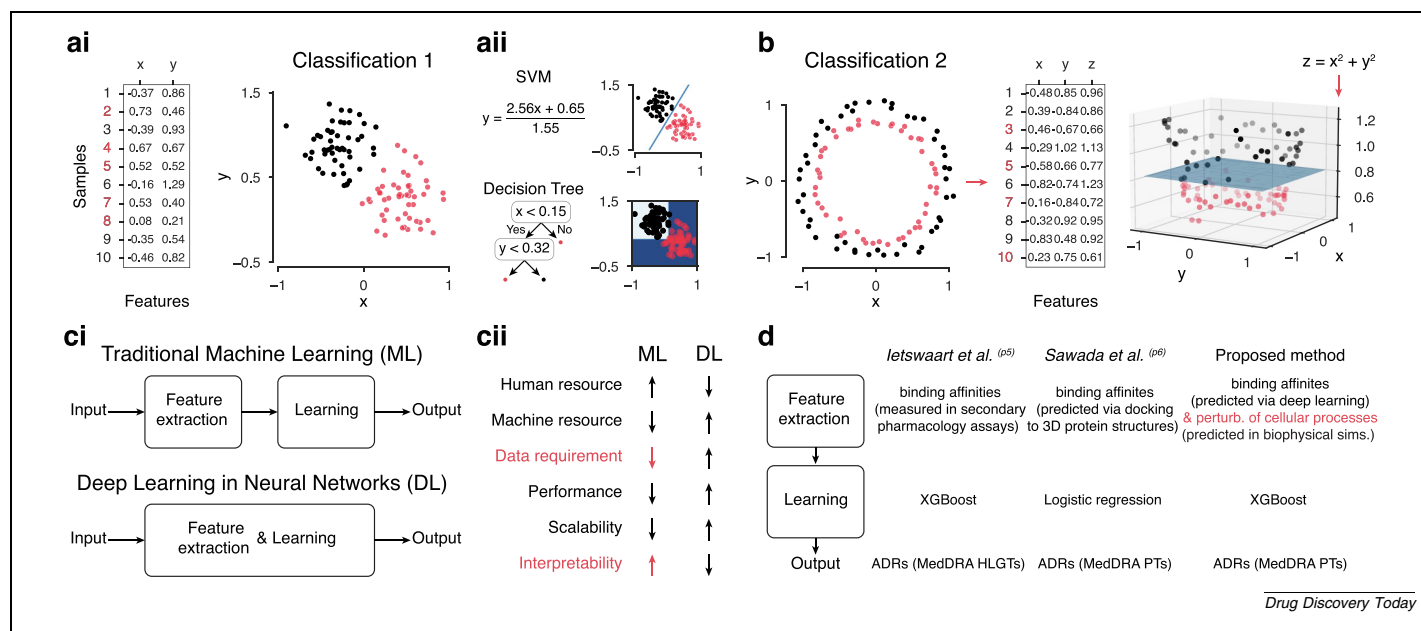


FIGURE 1

Overview of feature representation and machine learning models. **(ai)** First example of a 2D classification task (black vs red). Table on the left shows 10% of the data points in detail. Colors on the y axis match the class of the given data point. **(aai)** Two possible solutions of the task using the original features with either support vector machines (SVM) (on top) or a decision tree (on the bottom). **(b)** Second example of a 2D classification task, with a newly engineered feature (z as the sum of squares) which makes it solvable by a SVM in 3D. **(c)** Overview of simple models using feature engineering and neural networks that build their own feature representation. **(d)** Comparison of key elements of two recent models with our proposed one. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Training simple models

Here, we present results obtained by training XGBoost models with the same codebase on the same side-effect data (59 803 drug–side-effect pairs from OnSIDES,^(p19) with 1069 unique drugs and 561 unique side effects as MedDRA^(p20) preferred terms). The purpose of this exercise is to showcase how the feature representations introduced above compare with traditional features like ATC codes and drug fingerprints in terms of performance. Owing to the imbalanced occurrence of side effects (with only a few drugs causing specific side effects), we trained separate models for each side effect to balance drugs across the cross-validation folds. Although training separate models is common practice in the field,^{(p2),(p3),(p6),(p7)} using balanced accuracy (defined as the mean of sensitivity and specificity) as the primary evaluation metric to address the imbalance is new. The models trained on ATC level 1 codes are superior in sensitivity (also known as recall), whereas the ones on Morgan fingerprints are superior in specificity (Figure 2c). The models trained on our newly proposed features (labeled as Cytocast on Figure 2c) balance these two, and outperform the others in terms of precision.

The models were further challenged in a real-world setup, using compounds that were not part of the training set. Idiopathic pulmonary fibrosis (IPF) is a progressive, chronic interstitial lung disease of unknown etiology and has a median survival prognosis of 3–5 years, currently treated with nintedanib and pirfenidone^(p21) (Figure 2d). Clinical trials and pharmacovigilance

data mining highlighted slightly different side-effect profiles of the two drugs.^{(p21),(p22)} Because they share the same ATC level 1 code (L), the model trained on those cannot differentiate between them and predicts the same 301 side effects for both (Figure 3e, left). Furthermore, according to the OnSIDES database^(p19) there are only 12 and 30 ADRs associated with the compounds (Figure 2d, left: dashed lines). Accordingly, the model based on level 1 ATC codes vastly overpredicts, which also explains its superior recall (also known as true-positive rate). Albeit both compounds are effective in slowing down disease progression, neither can reverse its course, thus finding new compounds to treat IPF is an active area of research. Rentosertib was recently discovered by AI and is already in a Phase II clinical trial^(p23) (Figure 2d). Because it does not yet have an ATC code, no predictions can be made with this model (Figure 2d, left). Although the problems could be mitigated by predicting even the ATC code and adding more levels to the feature set, it illustrates the problems of learning on not readily available features and predicting ADRs that are general representatives of the given drug category, and not specific to the compound itself. Concurrently, the models based on Morgan fingerprints can be used in the preclinical phase and are able to differentiate between the compounds – the interpretation of their predictions would only highlight similarities in the structures themselves (Figure 2d, bottom right). Furthermore, they not only fall short of the ones using bio-realistic features in balanced accuracy of the side effects

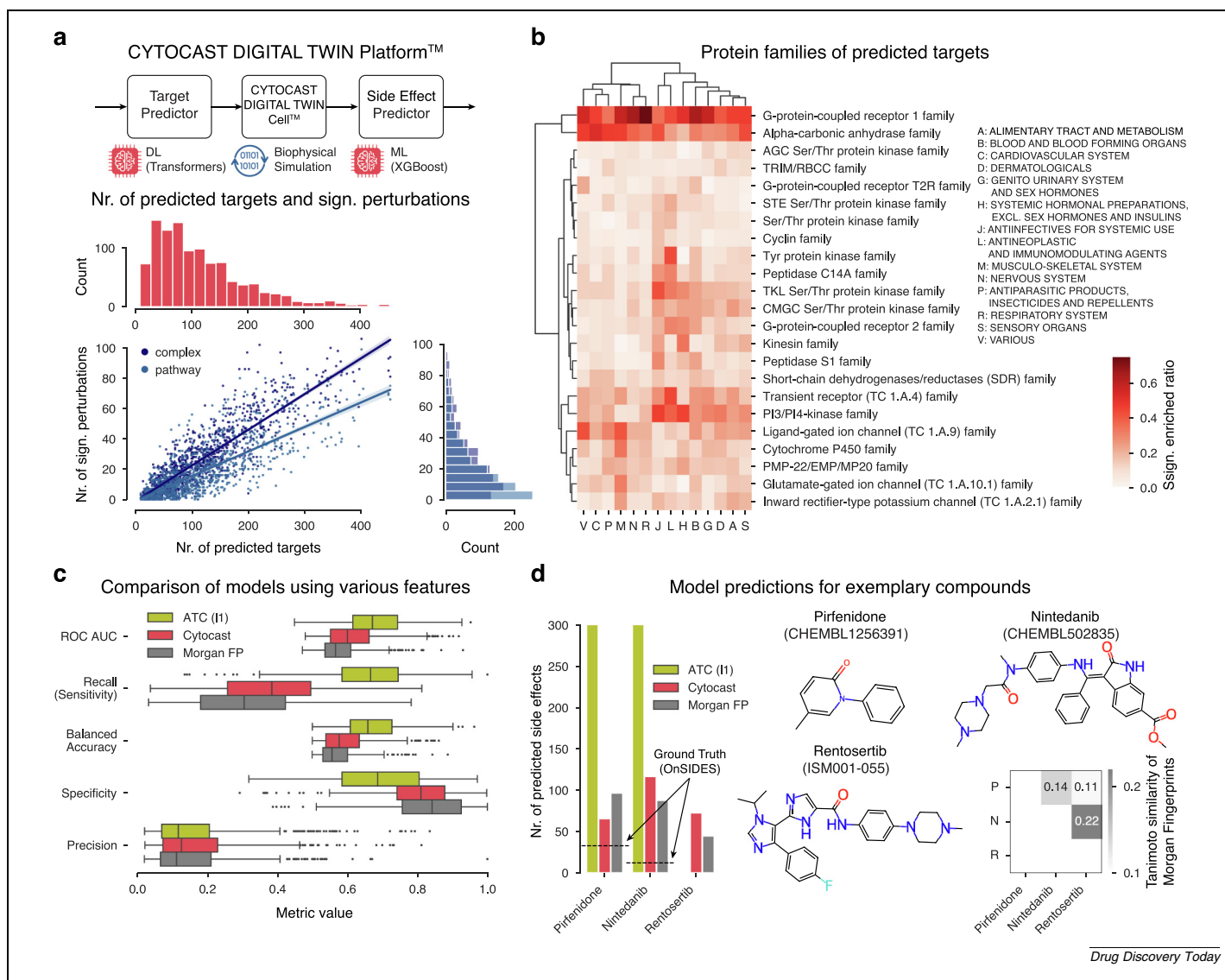


FIGURE 2

Models trained on various features. **(a)** Overview of Cytocast's pipeline on top and simulation derived features on the bottom: number of significantly perturbed complexes and pathways (775 and 206 unique ones, respectively) as a function of the number of predicted targets (4150 unique off-targets for 1069 training drugs). **(b)** Enrichment of protein families of predicted targets. Heatmap shows the ratio of drugs from the anatomical therapeutic chemical (ATC) category for which the given family was significantly enriched (Fisher's exact test, with significance threshold of 0.05). Only the most common families are shown. **(c)** Summary of various metrics of three different side-effect prediction models (same training labels, same models; but different features). **(d)** Comparison of model predictions for selected compounds. Number of predicted side effects on the left, structure of compounds (visualized using RDKit) and similarity of the compounds on the bottom right. On the left the ground truth side-effect count as well as the predictions based on ATC (level 1) codes for rentosertib are missing, because it is in active development.

(0.57 ± 0.06 vs 0.6 ± 0.08 , see Figure 2c) but also of the drugs (0.58 vs 0.65 and 0.48 vs 0.6 for pirfenidone and nintedanib, respectively).

Interpreting predictions

In this section we highlight the main strength of the proposed Cytocast model and show how actionable insight can be gained from interpreting its predictions. From the various available mathematical frameworks, we have chosen the Shapley Additive Explanations (SHAP) analysis, which distributes contribution among the features^(p24), thus helping to identify which off-target proteins, perturbed complexes and pathways lead to an

ADR being predicted. We observed that the contribution of the feature types in our training set varies across side effects (Figure 3a).

For the first example, we selected 'Stevens-Johnson syndrome', an off-target dominated side effect, which we correctly predicted for pirfenidone. By looking into the SHAP values of the potential off-target proteins we found that the prediction was 18% based on carbonic anhydrase 9 (Figure 3b), which was indeed a predicted off-target of pirfenidone. The same off-target was also the top contributor (8%) to a related true-positive prediction: toxic epidermal necrolysis. It is important to note here that this analysis does not necessarily describe the biological

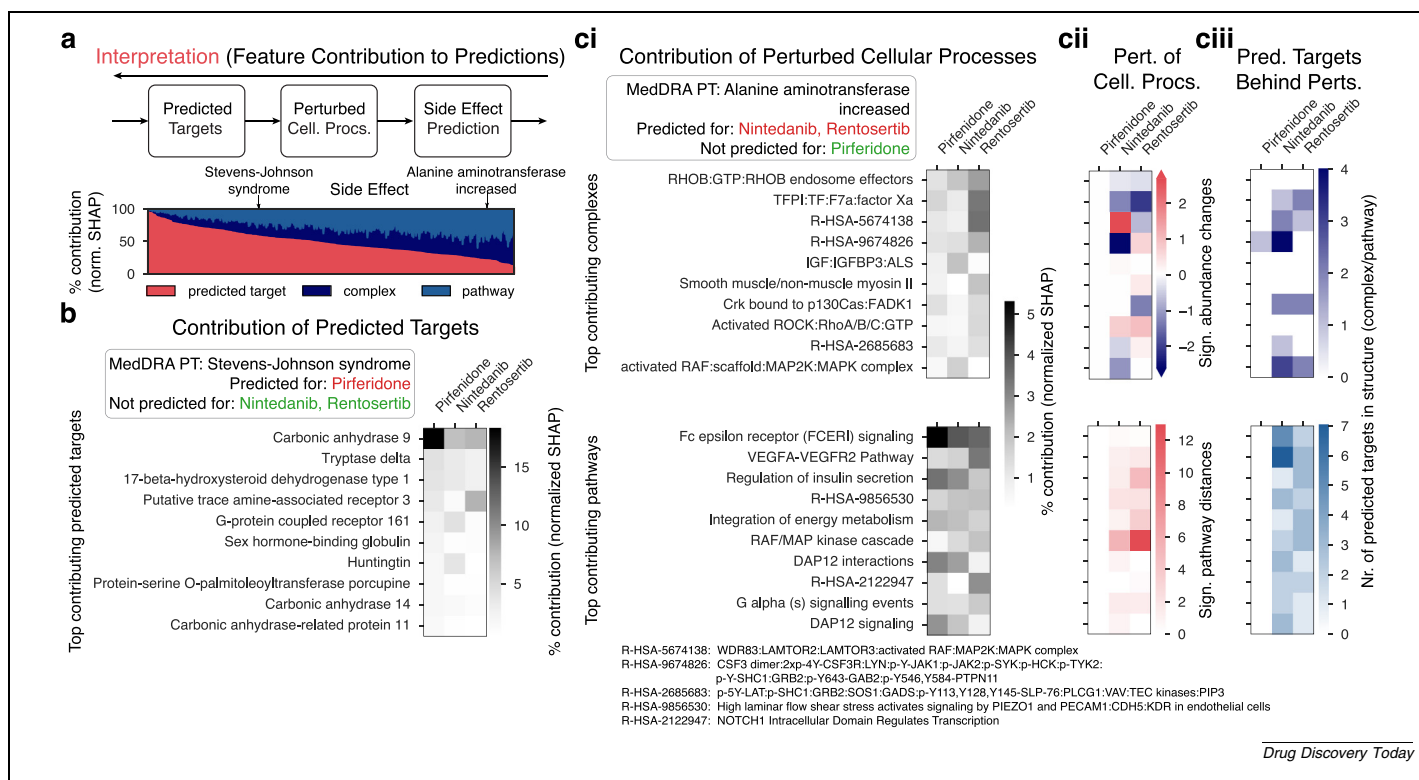


FIGURE 3

Interpretability of predictions. **(a)** Flowchart illustrating working backwards when interpreting the (Cytocast) model's predictions (see c for details) on top. Average contribution (normalized SHAP values) of the three different feature types in the training data on the bottom. Side effects (on the x axis) are sorted by the contribution of predicted targets. Arrows indicate the two side effects shown in detail on b, c. **(b)** First example of interpretability for a selected (predicted target dominated) side effect. The heatmap shows the contribution of the top ten predicted targets (selected based on and order by pirfenidone). **(c)** Second example of interpretability for a selected (perturbed complex and pathway dominated) side effect. **(ci)** The heatmaps show the contribution of the top ten complexes on top, and pathways on the bottom (selected based on and order by the sum of nintedanib and rentosertib). **(cii)** The heatmap on top shows the significantly (Mann–Whitney *U* test, with significance threshold of 0.001) perturbed complexes in the simulation. Positive/red hues indicate an increase, negative/blue hues a decrease in complex abundance, whereas white squares represent non-significant changes. The heatmap on the bottom shows the significantly perturbed signaling pathways in the simulation. Pathway distances are calculated as Euclidean distances of protein–protein interaction (PPI) count vectors. Significance test and missing values as for complexes above. **(ciii)** The heatmaps show the number of predicted targets that are part of the complexes on top and pathways on the bottom. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

mechanisms behind the ADRs, rather just showcases how our model understands them. However, in the case of carbonic anhydrase inhibitors there is a known link to severe side effects related to the skin such as the two above. ^(p25)

For our second example, we selected ‘alanine aminotransferase increased’, a side effect characterized by perturbations in complexes and pathways. This adverse event was also reported among the four most frequently occurring ADRs in the Phase IIa clinical trial of rentosertib by Xu and colleagues. ^(p23) Perturbations in the VEGFA/VEGFR2 signaling pathway are known to be associated with increased alanine aminotransferase levels, ^(p26) whereas some of the remaining pathways and complexes (especially those linked to insulin secretion and the MAPK pathway) are also intertwined with it (Figure 3ci). Most but not all of these were perturbed in the simulations for rentosertib (and nintedanib for which we also predicted it), which point out that features with value 0 can also heavily influence the predictions (Figure 3cii). These perturbed complexes and pathways can ultimately be linked back to off-target proteins (Figure 3ciii). Elimination of these off-targets, or even just lowering the compound

binding affinity to them, could be an actionable step against the ADR or, if that is not possible, it could help with patient stratification when preparing for the clinical trial. Furthermore, owing to the fully *in silico* nature of the study, these hypotheses about off-target reduction can be tested within the same framework.

Concluding remarks

In summary, whereas *in silico* tools are already in use for ADMET predictions ^(p27) and digital twin technology is revolutionizing healthcare in general, ^(p28) there is no industry standard tool that predicts ADRs. Some of the methods available in the literature appear to focus primarily on maximizing performance, ^(p29) neither addressing practitioner needs nor the availability of the data in inference time. The bigger problem is that of limited data. The most used side-effect resource to date is SIDER, ^(p30) which, owing to the lack of funding, was last updated >10 years ago. Many first-in-class drugs appear every year, and not taking these into account seriously limits the domain coverage of the trained models. Pharmacovigilance data, for example deposited to FAERS, ^(p31)

are also a valuable source of information. This applies even to drugs that have been on the market for a long time. For these reasons, we trained our models on side effects extracted by large language models (LLMs) from the latest drug labels and stored in the OnSIDES database.^(p19) Although being up to date, the LLM-based extraction also has its own error rate (Tanaka *et al.*^(p19) reported an F1 score of 0.875), and training models on labels that are not only limited and sparse but also noisy is a problem. Care must be taken in regularizing the models and correctly evaluating their performance in stratified cross-validation loops. Furthermore, although the most useful predictions would be about severe side effects that could put an end to clinical trials or lead to the withdrawal of marketed drugs, there are as of yet no publicly available data about ADRs that terminated clinical trials.

Acknowledging the paucity of available data, we strongly argue against training complex models, which, although being able to extract highly entangled relationships, need to see large amounts of samples. The added benefit of traditional ML techniques is that they are usually mathematically invertible or, if not, at least interpretable. To perform well, however, these models need good features, like measured^(p6) or predicted^(p7) off-target proteins. Our contribution lies in proposing to use even more of these biologically relevant features. Although one might argue that the proposed features from the digital twin cell add another layer of complexity to the model, we highlight that this feature representation is at least data-driven, not performance-driven, and can be traced back to unintentional bindings. This unique aspect of the proposed method could contribute to the extension or revision of secondary pharmacology assays, routinely run in drug discovery settings.^(p11) Furthermore, *in silico* screening of thousands of binders could also inform rational polypharmacology strategies.^(p32) Systematic comparison of predicted side-effect profiles could provide an additional layer of information to complement HTS and early hit selection as well as supplement decision making prior to candidate nomination and progression into

late-stage development and clinical evaluation. Not being a black box, this technology already provides actionable insights and in the future it might even aid the drug design process.

Declaration of interest

All authors are employees of Cytocast Hungary Kft., where AC-N is a board member as well. Furthermore, EF, IR and AC-N own equities or stocks of the company.

CRediT authorship contribution statement

András Ecker: Writing – review & editing, Writing – original draft, Visualization, Investigation, Formal analysis, Conceptualization. **Gergely Szabó:** Resources, Methodology. **János Szalma:** Software, Methodology. **Erzsébet Fichó:** Writing – review & editing, Formal analysis, Conceptualization. **István Reguly:** Writing – review & editing, Methodology, Conceptualization. **Attila Csikász-Nagy:** Writing – review & editing, Writing – original draft, Supervision, Conceptualization.

Funding

The project was, in part, supported by the ENFIELD Program under Contract: oc1-2024TIS-01-106. IR was further supported by the National Research, Development and Innovation Fund of Hungary (FK 145931), under the FK_23 funding scheme.

Acknowledgments

The authors thank all employees of Cytocast Hungary, especially advisor György M Keserű and commercial development leader Paul Lebeau for their helpful comments on earlier versions of the manuscript.

Data availability

The data supporting the findings of this study is available at <https://doi.org/10.5281/zenodo.18547374>.

References

- Sachdev K, Gupta MK. A comprehensive review of computational techniques for the prediction of drug side effects. *Drug Dev Res.* 2020;81:650–670.
- Toni E, Ayatollahi H, Abbaszadeh R, Fotuhi SA. Machine learning techniques for predicting drug-related side effects: a scoping review. *Pharmaceuticals.* 2024;17:795.
- Zhao H *et al.* Application of machine learning in drug side effect prediction: databases, methods, and challenges. *Front Comput Sci.* 2025;19, 195902.
- Jumper J *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature.* 2021;596:583–589.
- Galeano D, Li S, Gerstein M, Paccanaro A. Predicting the frequencies of drug side effects. *Nat Commun.* 2020;11:4575.
- Ietswaart R *et al.* Machine learning guided association of adverse drug reactions with in vitro target-based pharmacology. *EBioMedicine.* 2020;57, 102837.
- Sawada R, Sakajiri Y, Shibata T, Yamanishi Y. Predicting therapeutic and side effects from drug binding affinities to human proteome structures. *iScience.* 2024;27, 110032.
- Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. *Nature.* 1986;323:533–536.
- Yamins DLK, DiCarlo JJ. Using goal-driven deep learning models to understand sensory cortex. *Nat Neurosci.* 2016;19:356–365.
- Campillos M *et al.* Drug target identification using side-effect similarity. *Science.* 2008;321:263–266.
- Whitebread S, Hamon J, Bojanic D, Urban L. Keynote review: in vitro safety pharmacology profiling: an essential tool for successful drug development. *Drug Discov Today.* 2005;10:1421–1433.
- Rizzetto S *et al.* Context-dependent prediction of protein complexes by SiComPre. *npj Syst Biol Appl.* 2018;4:37.
- Wang M *et al.* Version 4.0 of PaxDb: protein abundance data, integrated across model organisms, tissues, and cell-lines. *Proteomics.* 2015;15: 3163–3168.
- Mistry J *et al.* Pfam: the protein families database in 2021. *Nucleic Acids Res.* 2021;49:D412–D419.
- Szklarczyk D *et al.* The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Res.* 2023;51:D638–D646.
- Mosca R *et al.* 3did: a catalog of domain-based interactions of known three-dimensional structure. *Nucleic Acids Res.* 2014;42:D374–D379.
- Balu S *et al.* Complex portal 2025: predicted human complexes and enhanced visualisation tools for the comparison of orthologous and paralogous complexes. *Nucleic Acids Res.* 2025;53:D644–D650.
- Ragueneau E *et al.* The reactome knowledgebase 2026. *Nucleic Acids Res.* 2026;54: D673–D681.
- Tanaka Y *et al.* OnSIDES database: extracting adverse drug events from drug labels using natural language processing models. *J Med.* 2025;6, 100642.
- Fescharek R *et al.* Medical Dictionary for Regulatory Activities (MedDRA). *Int J Pharm Med.* 2004;18:259–269.
- Man RK *et al.* A comparison of the effectiveness of nintedanib and pirfenidone in treating idiopathic pulmonary fibrosis: a systematic review. *Cureus.* 2024;16, e54268.

22. Sun X et al. Comparison of the safety profiles for pirfenidone and nintedanib: a disproportionality analysis of the US food and drug administration adverse event reporting system. *Front Pharmacol.* 2024;15, 1256649.
23. Xu Z et al. A generative AI-discovered TNIK inhibitor for idiopathic pulmonary fibrosis: a randomized phase 2a trial. *Nat Med.* 2025;31:2602–2610.
24. Lundberg S, Lee S-I. A unified approach to interpreting model predictions. *Proceedings 31st International Conference on Neural Information Processing Systems*, 2017:4768–4777.
25. Trubiano JA, Ostrov DA, Phillips EJ. Stevens-Johnson syndrome and toxic epidermal necrolysis associated with carbonic anhydrase inhibitors: epidemiology, genetics, and insights into mechanisms. *J Allergy Clin Immunol Pract.* 2019;7:2854–2856.
26. Fuentealba C et al. Evaluation of the effects of a VEGFR-2 inhibitor compound on alanine aminotransferase gene expression and enzymatic activity in the rat liver. *Comp Hepatol.* 2011;10:8.
27. Lee H, Kim J, Kim J-W, Lee Y. Recent advances in AI-based toxicity prediction for drug discovery. *Front Chem.* 2025;13, 1632046.
28. Katsoulakis E et al. Digital twins for health: a scoping review. *NPJ Digit Med.* 2024;7:77.
29. White N, Parsons R, Collins G, Barnett A. Evidence of questionable research practices in clinical prediction models. *BMC Med.* 2023;21:339.
30. Kuhn M, Letunic I, Jensen LJ, Bork P. The SIDER database of drugs and side effects. *Nucleic Acids Res.* 2016;44:D1075–D1079.
31. Potter E, Reyes M, Naples J, Dal Pan G. FDA Adverse Event Reporting System (FAERS) essentials: a guide to understanding, applying, and interpreting adverse event data reported to FAERS. *Clin Pharmacol Ther.* 2025;118:567–582.
32. Badkul A, Xie L, Zhang S, Xie L. Multimodal out-of-distribution individual uncertainty quantification enhances binding affinity prediction for polypharmacology. *Nat Mach Intell.* 2025;7:1985–1995.