

RESEARCH ARTICLE

Cite this: *RSC Med. Chem.*, 2024, 15, 1015

Deconvoluting low yield from weak potency in direct-to-biology workflows with machine learning†

William McCorkindale,^a Mihajlo Filep,^b Nir London,^b Alpha A. Lee^a and Emma King-Smith^{b*}

High throughput and rapid biological evaluation of small molecules is an essential factor in drug discovery and development. Direct-to-biology (D2B), whereby compound purification is foregone, has emerged as a viable technique in time efficient screening, specifically for PROTAC design and biological evaluation. However, one notable limitation is the prerequisite of high yielding reactions to ensure the desired compound is indeed the compound responsible for biological activity. Herein, we report a machine learning based yield-assay deconfounder capable of deconvoluting low yield from low potency to identify false negatives. We validated this approach by identifying promising SARS-CoV-2 main protease inhibitors with nanomolar activity that rivaled potency observed from the standard D2B workflow. Furthermore, we show how our framework can be utilized in a broad, *in silico* screen to produce compounds of similar potency as a D2B assay.

Received 16th December 2023,
Accepted 12th February 2024

DOI: 10.1039/d3md00719g

rsc.li/medchem

Introduction

Direct-to-biology (D2B) in combination with high throughput screening (HTS) streamlines the target discovery pipeline by obviating the need for separation and purification. Reactive chemical fragment and PROTAC design have both significantly benefitted from this style of HTS.^{1–4} However, D2B currently requires some careful experimental design to limit the noise of crude reactions. For example, reagent scavengers may be used to remove unwanted reaction components *via* filtration, but their application is limited to select reagents. For common chemistry such as peptide bond formation reactions, resins to remove residual coupling agent are commercially available, but this is not universal for all useful reactions.¹ Alternatively, crude reaction mixtures can be used for biological investigation, however, the reagents utilized must not interfere with the assay to form false negatives or false positives.³ Likewise for any byproducts formed, reaction optimization should be performed to limit the number of side products.² Often only higher yielding reactions are used in assays; lower yielding reactions are not

further investigated due to the confounding effects of lower yielding reactions in the biological assay. A robust method to deconvolute the inherent noise of the crude reaction from the biological activity would increase the scope of high throughput D2B to a more diverse range of chemistries.

Machine learning (ML) methods have been successfully applied to predicting a variety of medicinal chemistry relevant properties including QSAR, IC₅₀ prediction of purified compounds, and improved docking studies.^{5–7} The principal driving force behind these computational algorithms is identifying and subsequently modelling correlations between chemical or biological features and measured outcome. Our hypothesis is that a machine learning approach can effectively act as a denoising algorithm – picking out average trends in structure–activity relationships across chemical space amid noise such as variations in yield. Prior reports utilizing ML to identify data outliers, including false negatives, have encompassed a variety of approaches in the chemical and biological areas. Clustering algorithms such as principal component analysis (PCA), *k*-nearest neighbors (KNN), or hierarchical clustering are often employed.^{8–11} These methods allow for visual identification of outliers, distinguished by their literal graphical position of lying outside of the algorithmically-formed, dense cluster centers. Analytical chemistry and metabolomics have also employed the use of robust regression and classification models such as support vector machines (SVMs) and Adaboost to not only accurately model noisy systems but also to identify these deviating

^a Cavendish Laboratory, University of Cambridge, UK^b Department of Chemical and Structural Biology, The Weizmann Institute of Science, Israel^c Yusuf Hamied Department of Chemistry, University of Cambridge, UK.

E-mail: esk34@cam.ac.uk

† Electronic supplementary information (ESI) available. See DOI: <https://doi.org/10.1039/d3md00719g>

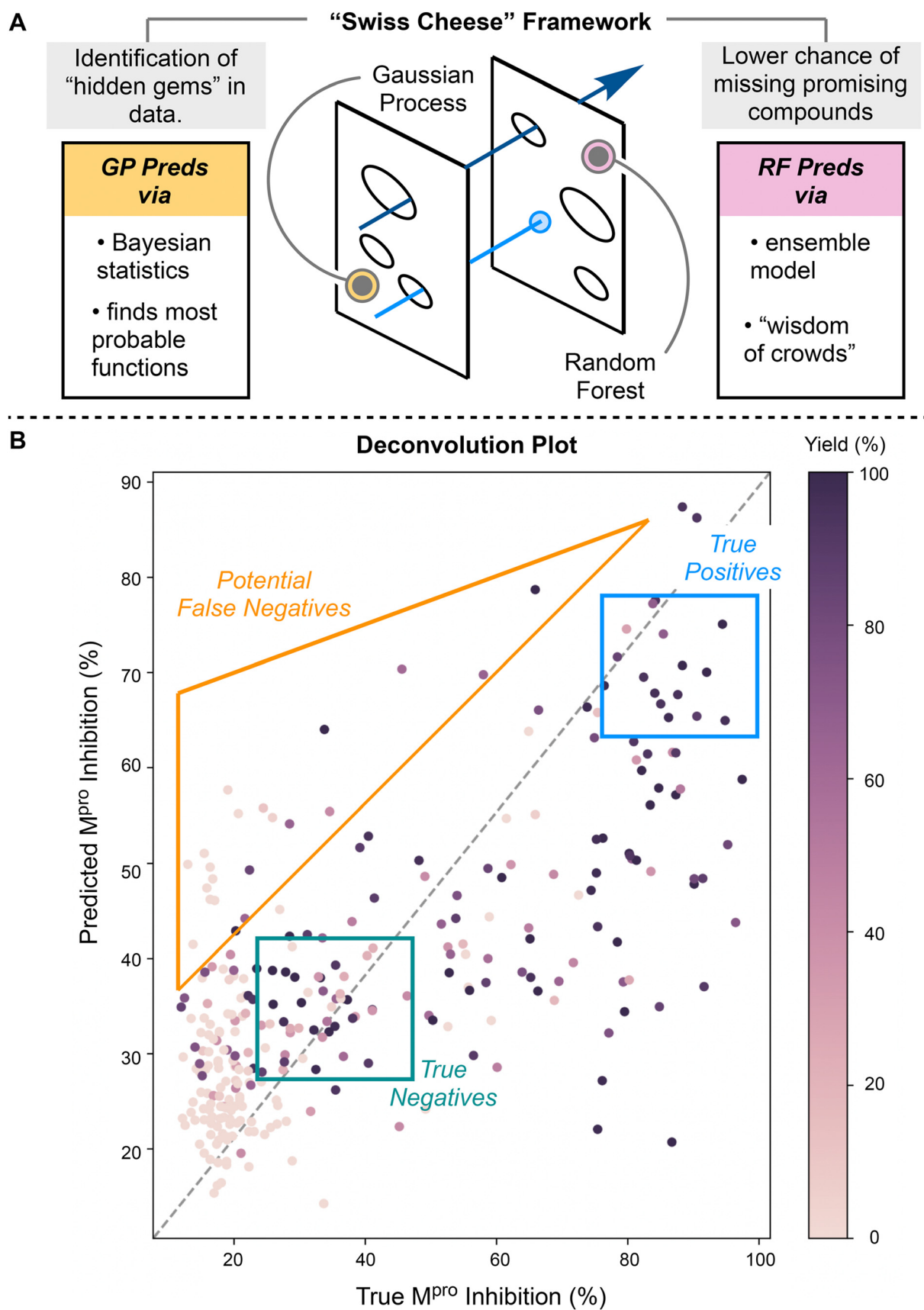


Fig. 1 Overview of ML workflow. (A) Graphical representation of the “Swiss cheese”. Hits missed by one model could be caught by the other. (B) Analysis of a deconvolution plot. Points inside the blue box are true positives (D2B hits). Points inside the teal box are true negatives. Points inside the orange triangle are potential false negatives.

datapoints.^{12,13} However, these detection methods typically have two defined end points. The first is to validate the robustness of their data as low outlier levels is indicative of high-quality data. The second is to treat outliers as unwanted noise and proceed with outlier removal for clearer modelling and prospective assessment. We believe that outliers, specifically false negatives, can be harnessed for further exploration when used in combination with high throughput D2B. The false negatives of a screen are molecules which, due to experimental noise, were deemed poorly performing in their associated assay, but in reality, are actually promising lead compounds. Herein, we report a “Swiss cheese” ML approach that utilizes two distinct ML paradigms to model confounded crude reaction mixture assay data (Fig. 1). Their combined predictions allowed for identification of false negative molecules that would have been missed in a traditional D2B workflow. Additionally, this paradigm can be put towards high throughput *in silico* screening of vast molecular space to identify new molecular scaffolds.

Results & discussion

For this investigation, we targeted the SARS-CoV-2 main protease, M^{Pro}, a non-structural protein that plays a key role in viral replication. M^{Pro} has been noted by several groups to be a promising target for drug design due to its importance in SARS-CoV-2's lifecycle and lack of related homologues in humans which has resulted in the success of approved SARS-CoV-2 antiviral treatments, nirmatrevir (Paxlovid) and ensitrevir (Xocova).^{14,15} Numerous high throughput assays to quantitatively measure M^{Pro} inhibition for rapid fragment screening have been developed.¹⁶ We opted for a previously reported fluorogenic assay which utilizes the synthetic fluorogenic M^{Pro} substrate, [5-FAM]-AVLQSGFR-[Lys(Dabcyl)]-K-amide.¹⁷

Prior effort from the COVID Moonshot consortium resulted in the rapid discovery of an orally bioavailable, nanomolar inhibitor of M^{Pro}, **1**, through the use of crowdsourcing, ML-directed synthetic complexity scores, and computationally guided structure activity relationship (SAR) screening. The authors noted the importance of the isoquinoline motif for M^{Pro} P1 pocket activity, but a high tolerance of structural variation of the M^{Pro} P2 pocket binding motifs. This allowed for the simplification of the original chromane ring to **1**'s chloro-tetrahydroisoquinoline, which furnished an N-centered handle for further SAR. Library-based derivatization of this handle with Schotten-Baumann sulfamidation resulted in a promising M^{Pro} inhibitor drug candidate with improved enzymatic inhibition, antiviral activity, and improved oral bioavailability, highlighting the importance of this tail.¹⁷ We began a further investigation upon the N-centered handle through the formation of similarly stable amide bonds with the fast turnaround of D2B screening, but including low yielding reactions that would be traditionally discarded. We

hypothesized that ML-guided hit identification would be able to identify promising scaffolds even in the noisy environment of low yielding, crude reaction mixtures. Low yielding but potent compounds will still register as low activity molecules. To this end, we synthesized a library of derivatives *via* EDC mediated peptide bond couplings (Fig. 1B). Given high throughput D2B is known for peptide bond formations, we believed this chemistry to be a good test-case for our deconfounding strategy. Thus, 300 amines were chosen, coupled to the commercially available M^{Pro} P1, P2 binder, acid **2**, and subjected without further purification or separation to fluorogenic M^{Pro} assay.

To identify these false negatives, we designed a “Swiss cheese” ML-based deconfounder. This was achieved by selecting two distinct ML techniques known to be highly accurate in molecular property prediction but orthogonal in their methods of generating predictions. The rationale was that ML techniques that model with different algorithms would result in a robust framework more accurate than if just one or the other was used. Hits missed by one algorithm still had the potential to be identified by the other. Agreement of predictions from both models indicated a high likelihood of experimental reality. A Random Forest regressor and Gaussian Process regressor were chosen as the two ML “Swiss cheese” slices (Fig. 1A). A Random Forest is an ensemble model technique formulated from a collection of decision trees.¹⁸ They encapsulate the “wisdom of crowds” where a single individual decision tree may be inaccurate, but their aggregate predictions are quite accurate. This mimics the well-known phenomenon that individuals can be inaccurate at guessing the correct answer, but the collective group has remarkable accuracy. Accurate prediction of compound solubility, quantitative structure–activity relationships (QSAR), and toxicity have been achieved with Random Forest modelling.^{19–21} Contrastingly, Gaussian Processes are based upon Bayesian statistics which predicts a distribution of functions whose sole restriction is that they pass through the known (experimental) datapoints. The properties of these functions are determined *via* the kernel, which characterizes the smoothness, periodicity, and shape of the underlying family of functions. Gaussian Processes have the additional benefit of predicting value and uncertainty. Intuitively, high uncertainty of a predicted value occurs when, at that given input, the range of values from the possible functions (determined from the kernel and known data points) is high. Gaussian Processes have accurately predicted lipophilicity, ADMET properties, and PROTAC potency in prior reports.^{6,22,23} Both models are non-parametric, making them ideal ML models for our “Swiss cheese” framework.

Whilst ML-based regression models that predict IC₅₀ values from pure substrate have been validated numerous times previously, we believe these methods are less suitable in the context of D2B.^{20,24–26} To develop any model that predicts the IC₅₀ outcomes (values from regression, or active/inactive from classification) of pure substrate, pure substrate must first be obtained.²⁷ However, two key features of D2B

are 1) the utilization of crude, unchromatographed compound in biological assays and 2) the ability to perform ultra-high throughput screens on nanogram quantities of substrate.⁴ Both of these key features make the traditional ML IC_{50} prediction platform incompatible. It would necessitate several hundred compounds be made on larger scales to ensure enough material for isolation and subsequent IC_{50} evaluation; a slower and less high throughput method which runs contrary to the goals of D2B. Thus, we propose that our model, with its capacity for false negative detection as a consequence of training upon crude IC_{50} data, lends itself as a better fit into D2B campaigns.

Deconvolution for false negative identification

We first began by identifying false negatives, these so-called “hidden gems”. These are compounds which, by nature of their low yield, have been labeled with low inhibitory activity. In a traditional D2B workflow, these compounds would be either a) excluded prior to the assay screening due to their low yields or b) require some level of scavenger-assisted separation to normalize the concentration of the potentially active compounds. Gaussian Process and Random Forest regressors were individually trained on the experimental M^{pro} inhibition data from the 300 crude amides (Fig. 2B, Table S1†). Due to the conserved nature of the acid, the Morgan fingerprint of the

amine moiety was used as the featurization of each compound. The predictions for both models, visually, lead to an easy identification of true positives and true negatives: molecules that have high yields and high or low activity, respectively. However, we can also identify potential false negatives (“hidden gems”): compounds that were determined by to be structurally represented in the high activity regime, but experimentally determined to have low yield and low inhibitory activity (Fig. 1B, S1 and S2†). The top 5 compounds that were determined to meet the aforementioned criteria (high predicted activity by both Gaussian Process and Random Forest, low experimental activity, low yield as determined by LC/MS) were selected as promising false negatives. These compounds, 3–7, had reported less than 35% M^{pro} inhibitory activity *via* the fluorogenic assay and a yield under 50%. Of the five compounds, 4 were able to be synthesized in suitable purity. Whilst compound 6 was indeed inactivate (deconvolution failed), compounds 4 and 5 showed high nanomolar to low micromolar inhibitory activity and 7 and boasted an IC_{50} value of 81 nM, ranking it within the top 20 most potent D2B hits (Fig. 3 and S3†). This result showcased the ability of the “Swiss cheese” framework to find new promising molecules that would be typically bypassed. Considering the ease of implementation and the low data requirement to run the models, we believe it is a low risk, high reward strategy to augment established D2B pipelines.

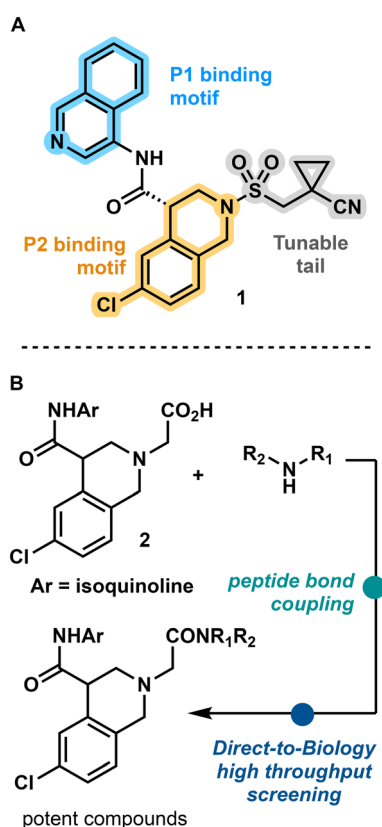


Fig. 2 Introduction to a new SARS-CoV-2 main protease (M^{pro}) inhibitor. (A) Structure of promising crowdsourced M^{pro} inhibitor. (B) Workflow for investigation of tunable tail fragment.

High throughput computational deconvolutional screening

Finally, we sought to investigate the potential of our “Swiss cheese” framework towards *in silico* screening. A set of 61 814 amines commercially available from Enamine were initially sourced and screened for structural alerts. Molecules with more than a single stereocenter were removed to avoid downstream complications in assessing potency. This resulted in 58 082 1° and 2° amines which were subjected to virtual screening by our trained “Swiss cheese” framework to identify possible amides with potent M^{pro} inhibition, irrespective of their potential coupling yield. The top 20 compounds predicted by the “Swiss cheese” to have high M^{pro} inhibition were synthesized and assessed (Fig. S4 and S5†). Of these compounds, 19 were successfully formed in suitable purity, and 7 of them had nanomolar inhibitory activity that ranked them in the 96th percentile of hits from the initial 300 amine screen. This resulted in an increase of sub 100 nanomolar inhibitors from 12 to 19, a 58% increase with minimal screening and computational cost.

Notably the structure of the top 2 most potent D2B compounds, 8 and 9, are remarkably different from the *in silico* screen's top two compounds, 13 and 14. Whilst 8 and 9 both contain aniline motifs, typically avoided due to their propensity to form reactive metabolites,²⁸ this deleterious functionality is absent in not just 13 and 14, but all the top 5 predicted most potent compounds. Comparing the top *non-aniline* containing D2B hits revealed that this brief *in silico* screen resulted in compounds with similar

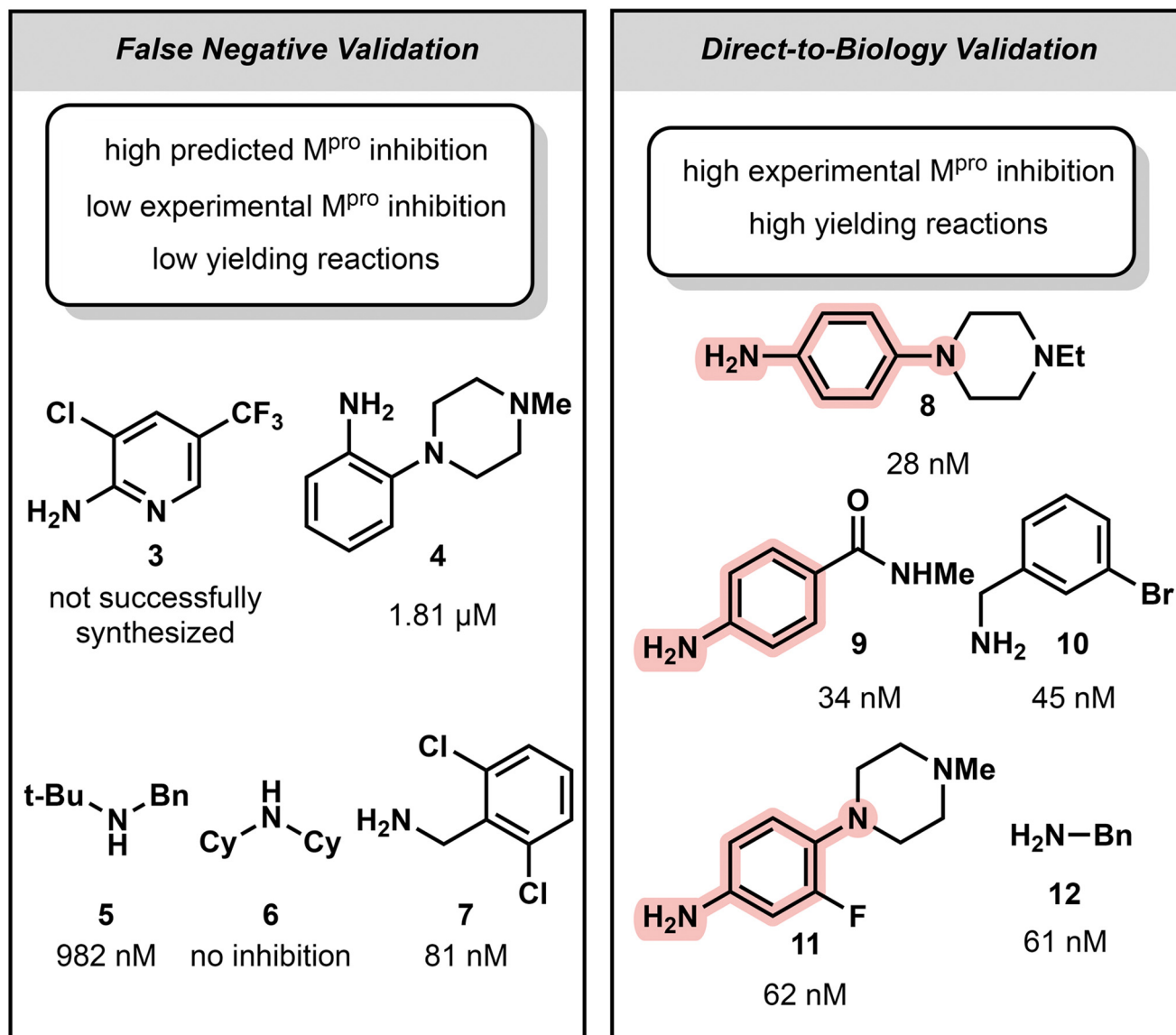


Fig. 3 Top validated amides (the amine used in the peptide bond forming reaction shown for clarity) of crude assay inhibition. IC_{50} values of the resulting amide shown below each compound. Selection criteria for compounds is listed. (Left panel) Top 5 putative and uncovered false negatives. (Right panel) Top 5 D2B hits from crude fluorogenic assay results (true positives). Deleterious aniline motif is highlighted in pink.

nanomolar M^{pro} inhibition, at 50 and 52 nM respectively, compared to 45 nM and 61 nM, highlighting this framework's capabilities not only in uncovering potent compounds that would have been discarded in a traditional D2B pipeline, but also as a valuable high throughput *in silico* screening module (Fig. 4).

Conclusion

We have showcased a dual machine learning framework that can expand the scope of D2B-compatible reactions by deconvoluting reaction yield and potency, allowing for low yielding reactions to be included. The “Swiss cheese” framework acted as a deconfounder of low yield from low M^{pro} inhibition and resulted in the elucidation of three active

compounds previously labeled as inactive. Additionally, we highlighted the framework's utility in a much broader screen of over 58 000 compounds; it would be challenging to validate such a number of compounds in a time- and cost-efficient manner. This *in silico* screen identified 7 new nanomolar potent compounds with valuable and non-deleterious structural motifs. We hope that future D2B and/or *in silico* screens can take advantage of this lightweight ML architecture which can be used in both low and high data environments to uncover new biologically valuable compounds and motifs.

Data availability

Code for this paper and the associated data can be found at: https://github.com/wjm41/deconvoluting_low_yield.

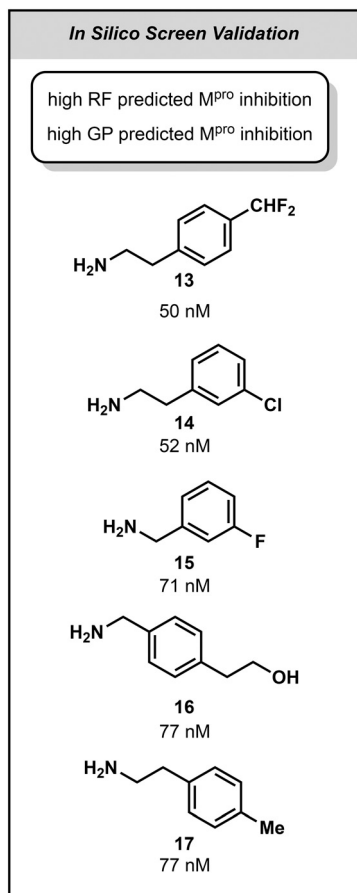


Fig. 4 The top 5 validated amides (the amine used in the peptide bond forming reaction is shown for clarity) from the “Swiss cheese” model of the *in silico* screen of 58 000+ Enamine amines. IC₅₀ values of the resulting amide shown below each compound.

Author contributions

WM and EKS performed the machine modelling. MF performed the biological assays. WM, NL, AAL, and EKS conceived of the work. EKS wrote the manuscript and all authors contributed to its editing.

Conflicts of interest

AAL is a co-founder and owns equity in PostEra Inc and Byterat Ltd. NL currently serves on the scientific advisory board of Monte Rosa Therapeutics, Larkspur Biosciences and Tesseract Medicines. WL, MF, and EKS declare no conflicts of interest.

Acknowledgements

We are grateful to insightful discussions with Lee group and Gaunt group members. WM acknowledges the support of the Gates Cambridge Trust. AAL acknowledges the Winton Programme for the Physics of Sustainability and the Royal Society (University Research Fellowship). NL would like to acknowledge funding from the Israel Science Foundation

(grant no. 3824/19), the Estate of Emile Mimran, Honey and Dr. Barry Sherman Lab, Dr. Barry Sherman Institute for Medicinal Chemistry and Nelson P. Sirotsky. EKS acknowledges funding from the Royal Society (Newton International Fellowship).

References

- C. E. Hendrick, J. R. Jorgensen, C. Chaudhry, I. I. Strambeanu, J.-F. Brazeau, J. Schiffer, Z. Shi, J. D. Venable and S. E. Wolkenberg, Direct-to-Biology Accelerates PROTAC Synthesis and the Evaluation of Linker Effects on Permeability and Degradation, *ACS Med. Chem. Lett.*, 2022, **13**, 1182–1190.
- R. Stevens, E. Bendito-Moll, D. J. Battersby, A. H. Miah, N. Wellaway, R. P. Law, P. Stacey, D. Klimaszewska, J. M. Macina, G. A. Burley and J. D. Harling, Integrated Direct-to-Biology Platform for the Nanoscale Synthesis and Biological Evaluation of PROTACs, *J. Med. Chem.*, 2023, **66**, 15437–15452.
- R. P. Thomas, R. E. Heap, F. Zappacosta, E. K. Grant, P. Pogány, S. Besley, D. J. Fallon, M. M. Hann, D. House and N. C. Tomkinson, A direct-to-biology high-throughput chemistry approach to reactive fragment screening, *Chem. Sci.*, 2021, **12**, 12098–12106.
- N. J. Gesmundo, B. Sauvagnat, P. J. Curran, M. P. Richards, C. L. Andrews, P. J. Dandliker and T. Cernak, Nanoscale synthesis and affinity ranking, *Nature*, 2018, **557**, 228–232.
- S. Kapsiani and B. J. Howlin, Random forest classification for predicting lifespan-extending chemical compounds, *Sci. Rep.*, 2021, **11**, 13812.
- O. Obrezanova, G. Csányi, J. M. R. Gola and M. D. Segall, Gaussian Processes: A Method for Automatic QSAR Modeling of ADME Properties, *J. Chem. Inf. Model.*, 2007, **47**, 1847–1857.
- C. Yang, E. A. Chen and Y. Zhang, Protein–ligand docking in the machine-learning era, *Molecules*, 2022, **27**, 4568.
- T. Schroeter, *Machine Learning in Drug Discovery and Drug Design*, 2009.
- L. B. Akella and D. DeCaprio, Cheminformatics approaches to analyze diversity in compound screening libraries, *Curr. Opin. Chem. Biol.*, 2010, **14**, 325–330.
- F. Plisson, O. Ramírez-Sánchez and C. Martínez-Hernández, Machine learning-guided discovery and design of non-hemolytic peptides, *Sci. Rep.*, 2020, **10**, 16581.
- A. H. Vo, T. R. Van Vleet, R. R. Gupta, M. J. Liguori and M. S. Rao, An overview of machine learning and big data for drug toxicity evaluation, *Chem. Res. Toxicol.*, 2019, **33**, 20–37.
- P. J. Rousseeuw, M. Debruyne, S. Engelen and M. Hubert, Robustness and Outlier Detection in Chemometrics, *Crit. Rev. Anal. Chem.*, 2006, **36**, 221–242.
- K. Chetnik, L. Petrick and G. Pandey, MetaClean: a machine learning-based classifier for reduced false positive peak detection in untargeted LC–MS metabolomics data, *Metabolomics*, 2020, **16**, 1–13.

- 14 Z. Jin, X. Du, Y. Xu, Y. Deng, M. Liu, Y. Zhao, B. Zhang, X. Li, L. Zhang and C. Peng, Structure of Mpro from SARS-CoV-2 and discovery of its inhibitors, *Nature*, 2020, **582**, 289–293.
- 15 A. Morris, W. McCorkindale, N. Drayman, J. D. Chodera, S. Tay, N. London and C. M. Consortium, Discovery of SARS-CoV-2 main protease inhibitors using a synthesis-directed de novo design model, *Chem. Commun.*, 2021, **57**, 5909–5912.
- 16 J. Ihssen, G. Faccio, C. Yao, T. Sirec and U. Spitz, Fluorogenic in vitro activity assay for the main protease Mpro from SARS-CoV-2 and its adaptation to the identification of inhibitors, *STAR Protoc.*, 2021, **2**, 100793.
- 17 M. L. Boby, D. Fearon, M. Ferla, M. Filep, L. Koekemoer, M. C. Robinson, The COVID Moonshot Consortium, J. D. Chodera, A. A. Lee, N. London, A. von Delft and F. von Delft, Open science discovery of potent noncovalent SARS-CoV-2 main protease inhibitors, *Science*, 2023, **382**, eabo7201.
- 18 K. Fawagreh, M. M. Gaber and E. Elyan, Random forests: from early developments to recent advancements, *Syst. Sci. Control. Eng.*, 2014, **2**, 602–609.
- 19 D. S. Palmer, N. M. O'Boyle, R. C. Glen and J. B. O. Mitchell, Random Forest Models To Predict Aqueous Solubility, *J. Chem. Inf. Model.*, 2007, **47**, 150–158.
- 20 V. Svetnik, A. Liaw, C. Tong, J. C. Culberson, R. P. Sheridan and B. P. Feuston, Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling, *J. Chem. Inf. Comput. Sci.*, 2003, **43**, 1947–1958.
- 21 A. Matsumoto, S. Aoki and H. Ohwada, Comparison of random forest and SVM for raw data in drug discovery: prediction of radiation protection and toxicity case study, *Int. J. Mach. Learn. Comput.*, 2016, **6**, 145.
- 22 T. S. Schroeter, A. Schwaighofer, S. Mika, A. Ter Laak, D. Suelzle, U. Ganzer, N. Heinrich and K.-R. Müller, Predicting Lipophilicity of Drug-Discovery Molecules using Gaussian Process Models, *ChemMedChem*, 2007, **2**, 1265–1267.
- 23 E. Semenova, M. L. Guerriero, B. Zhang, A. Hock, P. Hopcroft, G. Kadamur, A. M. Afzal and S. E. Lazic, Flexible Fitting of PROTAC Concentration–Response Curves with Change-point Gaussian Processes, *SLAS Discovery*, 2021, **26**, 1212–1224.
- 24 A. Park, M. Joo, K. Kim, W.-J. Son, G. Lim, J. Lee, J. H. Kim, D. H. Lee and S. Nam, A comprehensive evaluation of regression-based drug responsiveness prediction models, using cell viability inhibitory concentrations (IC50 values), *Bioinformatics*, 2022, **38**, 2810–2817.
- 25 H. R. Teles, L. L. Ferreira, M. Valli, F. Coelho and A. D. Andricopulo, Hierarchical Clustering and Target-Independent QSAR for Antileishmanial Oxazole and Oxadiazole Derivatives, *Int. J. Mol. Sci.*, 2022, **23**, 8898.
- 26 A. Qureshi, H. Tandon and M. Kumar, AVP-IC50Pred: Multiple machine learning techniques-based prediction of peptide antiviral activity in terms of half maximal inhibitory concentration (IC50), *Pept. Sci.*, 2015, **104**, 753–763.
- 27 A. P. Lind and P. C. Anderson, Predicting drug activity against cancer cells by random forest models based on minimal genomic information and chemical properties, *PLoS One*, 2019, **14**, e0219774.
- 28 A. F. Stepan, D. P. Walker, J. Bauman, D. A. Price, T. A. Baillie, A. S. Kalgutkar and M. D. Aleo, Structural alert/reactive metabolite concept as applied in medicinal chemistry to mitigate the risk of idiosyncratic drug toxicity: a perspective based on the critical examination of trends in the top 200 drugs marketed in the United States, *Chem. Res. Toxicol.*, 2011, **24**, 1345–1410.