scientific reports

Check for updates

OPEN

Multi-omics feature engineering driven by biomedical foundation models improves drug response prediction for inflammatory bowel disease patients

Laura-Jayne Gardiner[1]✉, Jennifer Kelly[1], Ashley Evans[1], Stephen Checkley[2], Karen Bingham[3], Graeme Macluskie[3] & David Bunton[3]

Using biomedical foundation models (FMs) for inference on small cohorts, represents a promising and practical route to advance drug response biomarker discovery and target identification. Here, we demonstrate this via an innovative data-driven inference workflow, using a fine-tuned, biomedical FM. We study multi-omics (genomic, transcriptomic) data and predict pharmacological responses, both from surgical diseased tissue of inflammatory bowel disease (IBD) patients. We use FM inference to inform feature selection and feature engineering strategies, where FM-derived features provide advantage for predicting IBD patient drug response and target identification. Firstly, calculating drug-target binding affinity (BA), enabling prioritisation of protein/gene targets and associated SNPs for drugs of interest. Secondly, using patient SNPs to mutate reference proteins and assess impact on drug BA. Thirdly, building strategies to fuse BAs and transcriptomics. Additionally, we created an open-source Model Context Protocol server, making our FM inference example accessible to the community via AI agents and natural language prompts.

Keywords Biomedical foundation models, Feature engineering, DNA, SNPs, Drug response, AI Agents

Foundation models (FMs) have revolutionised Artificial Intelligence (AI) and are now increasing its impact on real-world applications[1,2]. A FM is a large AI model trained on a vast quantity of unlabelled data. This technology exploits self-supervised training, where the AI model learns to reconstruct missing data pieces that are intentionally masked out. The result is an AI model that forms a solid base that can be used for customisation to specific downstream tasks i.e., fine-tuning tasks. Recent studies have shown the feasibility of applying FMs to the field of genetics profiled via omics. These studies highlight generalised pre-trained FMs that are built using sequences, such as DNA (DNABERT[3]), RNA (BMFM-RNA[4], scBERT[5], scGPT[6], Geneformer[7]) and protein sequence (AlphaFold 3[8], MAMMAL[9]), and models now extend beyond language models alone. These existing models can perform tasks such as predicting gene expression (transcriptomic response) from DNA sequence, predicting protein–protein and protein-drug interactions and TF/RNA/protein binding sites, cell type annotation, gene perturbation prediction, disease classification, drug response prediction and for functional classification of genes and transcripts.

For traditional AI and Machine Learning (ML) in the multi-omics domain, it is common for researchers to develop models trained on a single or small number of cohorts or studies. As such, there is a tendency for derived models to not generalise well to new datasets since they are limited to the problem and the characteristics of the data, on which they were trained. Large pre-trained models or FMs are a potential solution to this problem, and biomedical FMs offer a powerful approach to increase understanding of complex biological systems and disease mechanisms. However, the use of these technologies for real-world discovery applications is often hindered by compute availability, computational complexities, a lack of explainability, and the requirement for interdisciplinary expertise to build and train models. Although fine-tuning a model is significantly less computationally and resource intensive, the need for GPU and specialised expertise can similarly render this

[1]IBM Research Europe, The Hartree Centre, Sci-Tech Daresbury, Daresbury, Warrington WA44AD, UK. [2]STFC, The Hartree Centre, Sci-Tech Daresbury, Daresbury, Warrington WA44AD, UK. [3]REPROCELL Europe Ltd, West of Scotland Science Park, Glasgow G200XA, UK. ✉email: Laura-Jayne.Gardiner@ibm.com

option as inaccessible to researchers. This work addresses this need by exploiting existing, fine-tuned, open-source, multi-omics, biomedical FMs for inference, using derived inferences to gain insight directly and to engineer new features that we use in explainable ML workflows to predict patient drug response.

The FM that we use as a focal example is MAMMAL[9]—Molecular Aligned Multi-Modal Architecture and Language—a multi-task FM that has been pre-trained on large-scale biological datasets across diverse modalities, including proteins and small-molecules. MAMMAL has been fine-tuned and evaluated on eleven diverse downstream tasks, where it reaches a new state of the art (SOTA) in nine tasks. This includes a drug-protein or drug-target binding affinity (BA) calculation task which is at the heart of this study. Accurate prediction of drug-target BA is essential in the early stages of drug discovery. MAMMAL predicts BAs using pKd, the negative logarithm of the dissociation constant, which reflects the strength of the interaction between a small molecule (drug) and a protein (target). Although we use MAMMAL as a demonstration, our approach is not limited to MAMMAL, in fact any FM (or classic task-specific model) that effectively performs the prediction of drug-target BA from our desired inputs would be appropriate.

Our example application task aims to understand inter-patient variation in drug response. The ability to stratify patients based on their response to a specific drug is key to facilitating the recommendation of personalised interventions or treatments. Furthermore, during drug development, effective stratification of patients prior to clinical trials could help mitigate high failure rates during Phase II/III trials. Therefore, it is important to design accurate preclinical test systems and models to inform patient selection for clinical trials. Here, we focus on drug development for Inflammatory Bowel Diseases (IBDs). Almost 5% of the world's population, one in ten people in the UK, are affected by an autoimmune disease such as IBD[10]. Our cohort covers the two primary conditions: ulcerative colitis (UC) and Crohn's disease. Both are long-term inflammatory conditions of the gut, with UC confined to the colon (large intestine), and Crohn's disease affecting any part of the digestive system. These diseases are clinically, immunologically and genetically heterogeneous and result from complex host and environmental interactions. Currently there is no cure for either disease and people with IBD will likely require treatment throughout their lives. We know that drugs that target pathways implicated by genetics have a far higher chance of being effective[11], however, while over 240 IBD risk loci have been identified, fewer than 10 have been mechanistically resolved[12]. There is a remaining need to understand why certain medications are effective for different patients and to relate this to their demographic information and genetic makeup[13].

From ex vivo human diseased fresh tissues from IBD patients, REPROCELL Europe Ltd derived matched multi-omics (genomic, transcriptomic) data that we use to build models, and pharmacological responses (measured inflammatory cytokine TNFα level with/without test drug) as a proxy for preclinical drug efficacy that is our target to investigate/predict. TNFα is a proinflammatory cytokine that is a known mediator and disease driver in IBD, and in inflammatory diseases generally[14–16]. In this study we focus on a single test drug (prednisolone), a standard first line treatment for IBD, though our methods could be applied to any test drug. We mirrored the common 'all comers' clinical trial patient recruitment with only the donor inclusion criteria of a clinical diagnosis of UC or Crohn's, as confirmed by a clinical pathologist[17]. Therefore, the data sets and donor numbers in this study are representative of the scope of a typical preclinical drug development project using ex vivo human tissues, including 51 patients. Previous work assessed a small sub-cohort of this cohort, developing an explainable ML workflow combining genomic and demographic data to predict drug response to inform clinical trial and precision medicine strategies[18].

In this study, we firstly use FM drug-target BA inference to calculate BAs for reference proteins, enabling prioritisation of targets and associated genes/SNPs for a drug of interest. We secondly go on to model IBD pharmacological patient drug response primarily from genomics and transcriptomics data, while also identifying the most predictive features as part of a transparent ML precision medicine strategy. As part of this, we perform FM inference-driven feature engineering using patient SNPs (genomics) to infer impact on drug-target BA. We develop further feature engineering strategies to fuse FM-derived BAs with transcriptomics. Using explanation of our model's predictions we infer unique multi-modal combinations of important proteins, associated genes and their SNPs that are predictive of patient drug response. We showcase innovative *data driven* workflows where FMs can provide discernible advantage for the inference of patient drug response from multi-omics data. In combination these approaches can guide biological discovery for biomarker identification, target prioritisation, and can derive and rank new features for downstream predictions. Moreover, we use only fine-tuned FMs to showcase the benefits of existing open-source resources for small omics cohorts. Finally, by combining our open-sourced MCP server (hosting our FM inference task) with an AI agent (Claude, ChatGPT or Langflow etc.), we make complex aspects of our workflow accessible via natural language prompts. Specifically, we enable a user to perform FM inference to calculate drug-target BAs. In doing so, we expand FM accessibility for non-experts and for those with limited computational resource.

Results and discussion

Overall workflow

In Fig. 1 we show our overall workflow. This includes a Feature Selection section, comparing usage of standard approaches with FM inference for feature ranking i.e., enabling selection of targets for our drug of interest by calculating drug-target BA for reference proteins. In the Feature Engineering section of the workflow, we perform FM inference-driven feature engineering using patient SNPs to mutate reference proteins and measure impact on inferred drug-target BA. We then fuse FM-derived BAs with transcriptomics calculating an "Interaction State" and compare with traditional representations. We use these features to build ML models that predict patient pharmacological drug response in the "Predictive Modelling" section. Finally, using explanation of our model's predictions we infer unique multi-modal combinations of features that are predictive of patient drug response in the "Insights" section.

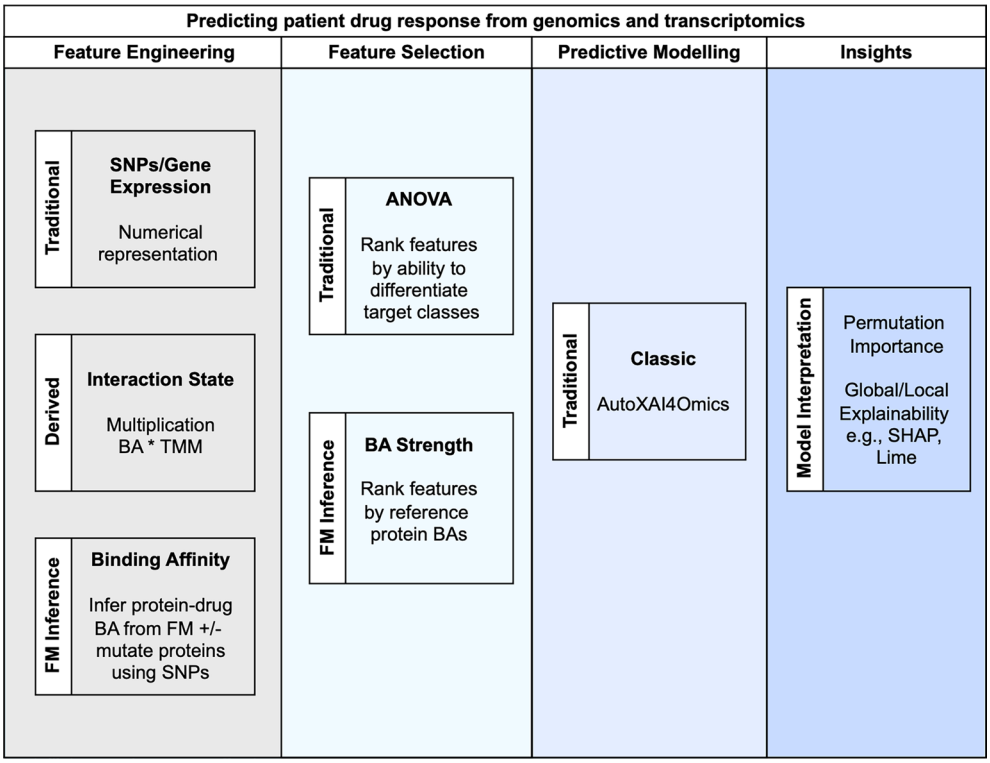

Fig. 1. Workflow summarisation – predicting patient drug response from genomics and transcriptomics. Outlining the key steps taken in this study to prepare features for ML (Feature Engineering), to decrease feature numbers (Feature Selection), to perform ML (Predictive Modelling) and to provide explainability of models (Insights). Generated features include SNPs, BAs (derived from reference and mutated proteins), gene expression values (TMM normalised) and interaction states (BA's x TMMs for a protein and associated gene).

FM-Inference driven feature selection

In this study we compared two approaches for Feature Selection (Fig. 1-Feature Selection). The first emulates classic feature selection i.e., finding a specified number of "best features" using feature ranking derived from ANOVA *F*-values for each column of features in the training data in relation to the target. The ANOVA F-test determines how well each feature separates the different classes in the target variable. ANOVA was selected as a widely used univariate baseline to illustrate the comparative effects of FM-based feature selection. While many multivariate methods exist, they require larger sample sizes or additional modelling assumptions that could obscure the specific contribution of FM inference. The second approach to feature selection that we designed also allows extraction of a specified number of "best features", but feature rankings are derived using FM-inference (Fig. 2a). Specifically, we rank proteins (and thus their associated genes and SNPs) using their calculated BAs with our drug of interest, prednisolone. BAs can be calculated in a data driven manner, with no reliance on knowledge of active sites or individual protein structures, using the FM MAMMAL that has been fine-tuned for calculation of drug-target BA. The expected inputs for the model are the amino acid sequence of the target and the SMILES representation of the drug. In this instance we use the SMILES representation of prednisolone and (in turn) calculate a BA for every unique reference protein amino acid sequence in the GRCh38 reference genome. We then order the reference proteins based on their BAs (a higher BA equating to a priority drug target).

The range of calculated BAs after FM inference varied from 4.65 to 7.61 (depicted in Fig. 2a). We focused on the top 5% of BAs i.e., the top 600 proteins when ranked in descending order by their BA with prednisolone. We then assessed if the predicted most strongly binding proteins appear to be biologically relevant. Prednisolone is a glucocorticoid that is used to treat inflammatory conditions. Reassuringly, the two main known targets for prednisolone NR3C1 – a glucocorticoid receptor and NR3C2 – a mineralocorticoid receptor, were found in the top 5% and top 1% of proteins respectively. We performed enrichment analysis for the 600 proteins (see Methods; Fig. 2b-c) and looking at the most enriched functional GO term categories, we see consistent themes relating to detection of a chemical stimulus and the G-protein coupled receptor signalling pathway. In this instance, chemical stimulus detection/perception is driven by a diverse set of olfactory receptors. The olfactory receptor (*OR*) genes have been found to be broadly expressed in a range of tissues including the colon[19], and are a category of G-protein-coupled receptors (GPCRs)—human membrane proteins that are an important class of drug targets. While steroid hormones were mainly thought to influence gene expression through nuclear receptors, evidence is now accumulating, which this work supports, for rapid, non-genomic effects of steroids, potentially via direct activation of GPCRs[20]. Synaptic transmission (cholinergic) is also an enriched category

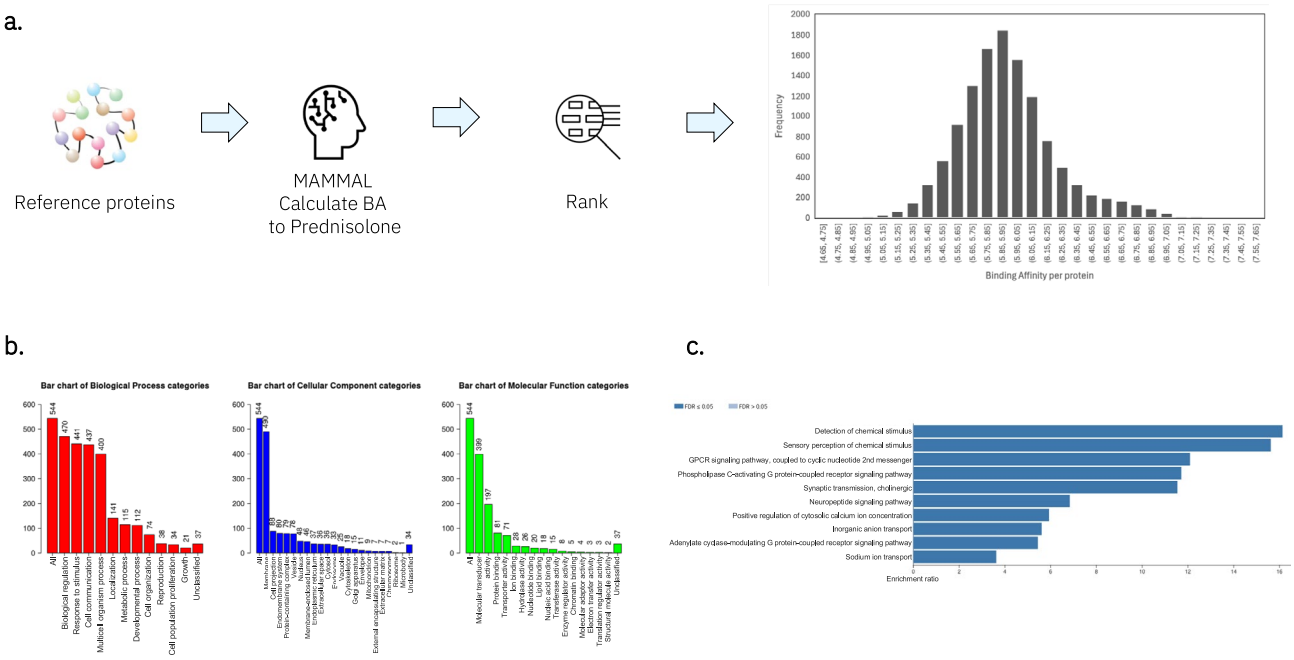

Fig. 2. Feature Selection using MAMMAL Inference. **(a)** Schematic detailing a high-level overview of the workflow used to derive feature rankings (at the protein level that we subsequently use to rank associated genes and SNPs) using BAs calculated for the drug prednisolone using the FM MAMMAL. After selection of the top 5% of proteins according to their BA with prednisolone; **(b)** Bar charts (see individual labels) illustrating the number of proteins that overlap annotated genes with the GO Slim terms from biological processes (red bar plot), cellular component (blue bar plot) and molecular function (green bar plot) ontologies respectively. **(c)** Bar chart to plot the enrichment results vertically with the bar width equal to the enrichment ratio from the gen set enrichment analysis.

in Fig. 2c, driven by a series of cholinergic receptors (muscarinic and nicotinic). Another corticosteroid that is similar to prednisolone, called dexamethasone, was shown to significantly increase the number of surface acetylcholine receptors (cholinergic receptors) on cultured human muscle[21].

When assessing the effectiveness of our feature selection, the combined evidence presented for selection of known targets and more widely the selection of biologically sensible enriched functional GO term categories, this appears to be an effective strategy. However, the best performance metric is to assess how predictive these features are when used directly in a ML model to predict patient drug response since this is our primary goal for this study.

FM-inference-driven feature engineering for genomics

We firstly focused on genomics data for the prediction of patient drug response, from which we derived SNPs across 51 patients (see Methods). The SNPs were used in two parallel analytic threads to derive features for downstream ML (Fig. 1-Feature Engineering). Firstly, as a representation of "raw" SNP data, we converted unique SNPs to either; 2 (homozygous alternate allele), 1 (heterozygous) or 0 (homozygous reference if no SNP was recorded for that patient at that position). Secondly, we used missense SNPs per patient i.e., those that result in a different amino acid being encoded in a protein, to mutate the 600 reference proteins that were identified during FM-driven feature selection. On average each patient had 11,378 missense SNPs, and across all the patients 2,666 unique missense SNPs fell within the gene space that overlapped the 600 reference proteins. BAs could then be calculated for each patient specific mutated set of 600 proteins using the MAMMAL and the SMILES representation of prednisolone. For each patient, and for each protein, the patient specific BA was compared to the matching reference protein BA (calculated by subtracting the reference BA from the patient BA so that a positive patient specific BA indicates an increased BA for the patient compared to that of the reference or unmutated protein). This second engineered feature set for ML thus included calculated patient specific BAs across the 600 proteins.

BAs improve predictions of drug response

Combining our feature selection and feature engineering approaches we collated three separate feature sets for downstream ML; (1) raw SNP data after traditional feature selection with ANOVA-F (Traditional Feature Selection + Traditional SNPs, 2666 features, which we denote as "Traditional-FS + SNPs"), (2) raw SNP data after feature selection for SNPs (2,666) overlapping the top 600 proteins with the strongest BAs to prednisolone derived using FM inference (FM-based Feature Selection + Traditional SNPs, 2666 features, which we denote as "FM-FS + SNPs") and (3) patient specific BAs per protein (derived using patient SNPs to mutate proteins and

compare to reference BAs) after feature selection for the top 600 proteins with the strongest BAs to prednisolone as derived using FM inference (FM-based Feature Selection + FM-based Patient-specific BAs, 2666 SNPs used across 600 proteins, which we denote as "FM-FS + FM-Encoded-SNPs"). Feature selection for set (1) was capped at 2,666 SNPs to match the data sets represented in feature sets (2) and (3). These three feature sets were then used to predict TNFa level i.e., our proxy for patient response to prednisolone, as a binary classification task where we predict high versus low TNFa level response groups using AutoXAI4Omics[22] (see Methods). As all human tissue organoculture pharmacological response data was normalised against a patient-matched non-treated control group, lower TNFα levels i.e., class 0, was interpreted as a better patient response to the drug.

Table 1 shows our results from ML with the three feature sets. We focus on the 'best' hyper tuned model for each feature set to make our initial comparison. The encoded algorithm within AutoXAI4Omics (see original publication for details[22]) suggests the 'best' or 'recommended' model for the user, comparing a range of seven classifiers based on a high accuracy coupled to a low performance variation across the different folds on cross-validation (CV) plus minimal overfitting observed when comparing training and test data performance (in this instance using F1-score) and prioritising balanced performance across the predicted classes. If we use traditional feature selection plus traditional SNP representations as our baseline (Traditional-FS + SNPs), comparing FM-based feature selection for traditional SNPs (FM-FS + SNPs), we saw a highly conserved F1 score on the held-out test data (overall test F1changed only by 0.005) and on CV. Although the test-set F1 values are similar between Traditional-FS + SNPs and FM-FS + SNPs, the FM-based features reduce overfitting, as indicated by a substantial drop in training F1 from 0.95 to 0.80 and a halving of CV variance. This suggests that FM-ranked proteins yield a more generalisable and stable feature set i.e., performing more consistently across our train and test datasets, even before additional FM-engineered features are included, while traditional ANOVA-based feature selection on the training data was more fitted to the training data. We noted particularly poor performance on CV for both feature sets (Table 1) with ranges of 0.34–0.61. This effect could also be amplified since our cohort is small. To ensure that optimisation artifacts were not driving performance differences, we used a single fixed classifier (Random Forest) to compare the performance of the FM-FS + SNPs to the Traditional-FS + SNPs features (Table S1). Here, we observed the same patterns of improved performance and reduced overfitting.

Next if we compare the effectiveness of using FM-derived patient-specific BAs instead of SNPs (FM-FS + FM-Encoded-SNPs) to Traditional-FS + SNPs and FM-FS + SNPs , we see a marked increase in performance for the held-out test data (overall test F1 increased by 0.09). In this instance the additional FM-driven feature engineering benefits ML performance on our unseen data. This effect was amplified through a marked increase in performance on CV (median F1 score increased by > 0.2). We did see some overfitting to the training data with this model, but at a lower level than for Traditional-FS + SNPs. Additionally, this test performance increase was echoed when we used a single fixed classifier (Random Forest) to compare feature sets (Table S1). In summary for FM-FS + FM-Encoded-SNPs, considering performance on the held-out test data (Figure S1a) and during CV and training (Figure S1b), the AutoKeras hyper tuned model was selected as 'best' by AutoXAI4Omics (Figure S1c). For this model ROC AUC values were consistently high across the classes (both 0.87) and during CV a median F1 score of 0.70 was observed with a standard deviation of 0.16 (Table 1). The confusion matrix for this model highlights only two samples that were incorrectly classified in the test set (Figure S1d).

While an accurate predictive model is desirable, when we are assessing small cohorts, our goal cannot be to produce a production-grade model that will be re-used on diverse populations downstream. Instead, our goal is to develop new ways to represent this data to interrogate our cohort and identify common genetic contributors to drug response. As such, it is imperative to understand the key drivers of model predictions i.e., the proteins that are most predictive in the model. SHAP is an explainability method that can be applied to any ML model.

Feature set	Train F1	Test F1	Overall train F1	Overall test F1	CV Med F1	CV F1 stdv	Classifier
Traditional-FS + SNPs	0.95, 0.95	0.73, 0.73	0.95	0.73	0.47	0.12	AutoXGBoost
FM-FS + SNPs	0.75, 0.84	0.67, 0.77	0.80	0.72	0.44	0.06	AutoSKLearn
FM-FS + FM-Encoded-SNPs	1.00, 1.00	0.83, 0.80	1.00	0.82	0.70	0.16	AutoKeras
FM-FS + TMM	1.00, 1.00	0.67, 0.60	1.00	0.63	0.64	0.15	RandomForest
FM-FS + FM-Interaction-State	1.00, 1.00	0.80, 0.83	1.00	0.82	0.62	0.26	AutoKeras
Integrated-FM-Features	1.00, 1.00	0.80, 0.83	1.00	0.82	0.77	0.15	AutoKeras

Table 1. ML performance across the tested feature sets. Shown here are the defined 'best' models per feature set. Feature sets: (Traditional-FS + SNPs) Traditional Feature Selection + Traditional SNPs, 2666 features, (FM-FS + SNPs) FM-based Feature Selection + Traditional SNPs, 2666 features, (FM-FS + FM-Encoded-SNPs) FM-based Feature Selection + FM-based Patient-specific BAs, 2666 SNPs used across 600 proteins. (FM-FS + TMM) Transcriptomic (bulk RNA-seq) TMM values for genes relating to the 600 proteins from FM-based Feature Selection, (FM-FS + FM-Interaction-State) Transcriptomic TMM values x FM-FS + FM-Encoded-SNPs values yielding "interaction states" for genes relating to the 600 proteins from FM-based Feature Selection, (Integrated-FM-Features) Concatenated feature set of individual feature sets (FM-FS + FM-Encoded-SNPs) + (FM-FS + TMM) + (FM-FS + FM-Interaction-State) using the AutoXAI4Omics ANOVA-based feature selection of 100 features. The performance metrics detailed include the per class F1 scores for the training dataset (Train F1) and the held-out test dataset (Test F1), the weighted average F1 score for the train (Overall Train F1) and held-out test data (Overall test F1), the median and standard deviation F1 on threefold cross validation (CV) and the classifier selected as producing the best hyper tuned model per feature set.

By combining game theory with local explanation, SHAP provides interpretations of how a model predicts a particular class for a given sample. Additionally, SHAP can rank important features for an ML model, explaining how each feature contributes positively or negatively to the prediction of a specific class. Figure S1e is a summary dot plot of the explainable ML results as computed by SHAP for our best model with FM-FS + FM-Encoded-SNPs, highlighting the top 15 features and their relative importance's for classification. In Figure S1e, a positive SHAP impact value of a feature for a sample represents a positive association with that feature value (patient specific BA) and the sample being assigned to a specific class, here class 0 is shown (patients assigned to class 0 indicate a lower TNFa level post-prednisolone treatment aka a better response).

SHAP analysis (Figure S1e) revealed that ten out of the top fifteen most influential features for predicting drug response were olfactory receptors (including OR4K15, OR8D2, OR5A2 and OR4C11 all in the top five). This is not surprising since during our previous enrichment analysis of the top 600 proteins (ranked by drug BA) the most enriched functional GO term categories were driven by members of the olfactory receptor (OR) family. For the most predictive feature from Figure S1e, an OR known as OR4K15, a higher prednisolone BA was an important predictor of good drug response. For protein OR4K15, 28 patients have one or more of seven SNPs (chr14:19,975,784 A/G, 19,975,853 A/T, 19,975,867 T/G, 19,976,234 T/C, 19,976,282 C/A, 19,976,429 T/C, 19,976,448 C/G) that change the protein sequence (Asn89Ser, Glu112Val, Ser117Ala, Val239Ala, Ala255Glu, Leu304Pro, Ile310Met). 23 patients have no SNPs, a further 3 patients have only chr14: 19,976,234 and an additional 3 patients have two SNPs (chr14: 19,975,784 and chr14:19,976,429), all of which have no effect on calculated BA (all patients show a baseline zero). The remaining 22 patients all have six SNPs, of which, four are not found in the baseline patients (chr14: 19,975,853, 19,975,867, 19,976,282, 19,976,448) and that decrease the BA of these patients to a conserved level. Thus, the SNPs drive a lower BA that in turn predicts higher inflammation levels post treatment. Patients with increased prednisolone BA are those with no (or fewer) SNPs and therefore expressing the unmutated (or less mutated) OR4K15 protein. This drives the prediction of class 0 or low TNFa (inflammation) and suggests that mutations in this protein may reduce a patient's response to prednisolone treatment. Olfactory regulators have been associated with regulation of intestinal inflammation and cell differentiation, and IBD patients often have decreased olfactory function[23]. From transcriptomic data that was generated from the same tissues (intestinal) that were used to generate the SNP data, we noted that across our cohort 263 OR genes were expressed. This supports the existing theory that OR genes are expressed in the colon and a role for ORs plausible.

Extending our feature set to include transcriptomics (a multi-omic view)

Next, we wanted to ascertain if adding another omics data layer would improve predictive capability. Transcriptomic data was generated from the same tissues, for the same 51 patients that we had pharmacological and genomic data for i.e., paired samples. Exploiting our FM-driven feature selection approach, we selected the transcriptomic features (TMM values) for those genes overlapping the top 600 proteins with the strongest BAs to prednisolone. This resulted in a fourth feature set for downstream ML (FM-FS + TMM) encompassing 325 genes across the 51 patients (since not all the 600 genes showed expression across the cohort). Table 1 shows our results from the best ML model for the prediction of patient prednisolone response or TNFa level with this feature set (including our best feature set so far– FM-FS + FM-Encoded-SNPs – as a comparison). The overall (and per-class) F1 score on the held-out test data was relatively low at 0.63, and overfitting was observed on the training data (F1 score 1.00), though notably performance on CV was highly similar to that of the test data (median F1 of 0.64), suggesting a degree of uniformity of performance on hold-outs across the cohort that we had not seen with previous feature sets.

We tested if the amalgamation of the transcriptomic data and the patient-specific BA would be useful i.e., could we concatenate the feature sets or adjust BAs according to gene expression level to incorporate both information sources into a single value per gene, per patient. As such, firstly we created a new feature set deriving what we named as "the interaction states" between our gene expression TMM values and patient-specific BAs for the 325 genes for which transcriptomic data was available (derived as TMM value x patient specific BA) – this is feature set FM-FS + FM-Interaction-State for our comparison. FM-FS + FM-Interaction-State performed better than FM-FS + TMM and similarly to our best feature set so far (FM-FS + FM-Encoded-SNPs) on the held-out test data and training data (Table 1). However, similarly to using TMMs directly, performance on CV was lower (median F1 score of 0.62) and highly variable (standard deviation 0.26).

We next considered if a concatenation of features might be worthwhile. From a biological viewpoint our feature sets FM-FS + FM-Encoded-SNPs, FM-FS + TMM and FM-FS + FM-Interaction-State might all independently add value to a model that predicts patient drug response. FM-FS + FM-Encoded-SNPs derives protein-drug BAs from patient genomic information and performs well generally in ML models. FM-FS + TMM includes genes whose expression level might be linked to drug response independently of binding of the gene's associated protein. If considered in addition, FM-FS + FM-Interaction-State could also help highlight those genes (whose derivative proteins) have strong BAs and are highly expressed or those that have poor BA's but aren't expressed. As a result, for comparison, we combined the patient specific BAs (FM-FS + FM-Encoded-SNPs) with the transcriptomic TMM values (for the 325 genes – FM-FS + TMM) and the interaction states (FM-FS + FM-Interaction-State) as a direct concatenation. This resulted in a feature set (Integrated-FM-Features) of 1230 expressions/interactions/BAs across the 51 patients. As this was a high dimensional dataset, we used the automated feature selection functionality from AutoXAI4Omics on the Integrated-FM-Features.

This combined feature set (Integrated-FM-Features), yielded our best performing model overall (built with AutoKeras), considering performance on the held-out test data (F1 score 0.82, Fig. 3a) and during training and CV (Fig. 3b). This hyper tuned model gave the best performance that we observed across all feature sets on CV (median F1 score of 0.77, range 0.59–0.88). The confusion matrix in Fig. 3c highlights only two samples in the test set that were incorrectly classified, similarly to Figure S1d for our previous best model with FM-FS + FM-

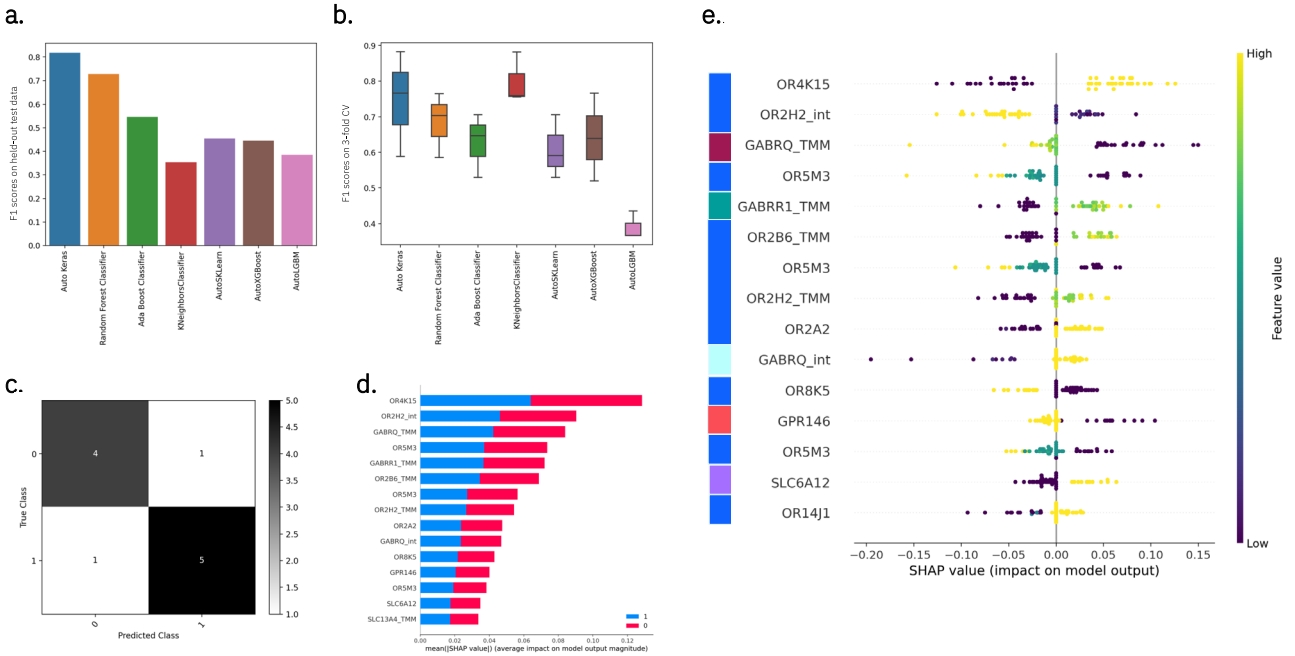

Fig. 3. Best performing ML model for Integrated-FM-Features. Evaluating the performance for our best performing ML model that was built with three concatenated feature sets; (FM-FS + FM-Encoded-SNPs) FM-based Feature Selection + FM-based Patient-specific BAs, (FM-FS + TMM) Transcriptomic TMM values for genes relating to the 600 proteins from FM-based Feature Selection, (FM-FS + FM-Interaction-State) TMM values x FM-FS + FM-Encoded-SNPs values for genes relating to the 600 proteins of interest. (**a**) Bar plot that shows the performance (F1 scores) of a range of hyper-tuned ML models for the prediction of the target for the held-out test data. (**b**) Box plots that show the performance (F1 scores) of a range of hyper-tuned ML models for the prediction of the target during cross validation. (**c**) The confusion matrix for the held-out test dataset for the best performing model (AutoKeras) and (**d**) summary SHAP bar plot showing how much each feature contributes to predictions on average across observations. The y-axis ranks the relative importance of the input features (protein names) on the best ML model—from high impact to low impact from top to bottom of the axis. (**e**) summary SHAP dot plot, the explainability output relating to the best performing model (AutoKeras) to predict class 0. SHAP global view of explanations for class 0 predictions. The y-axis of Figure S3e ranks the relative importance or of the input features on the best ML model—features are ranked from high impact to low impact from top to bottom of the y-axis and plotted against their SHAP impact values (x-axis). The colour bar groups proteins by superfamily. Rows are features (BAs per protein – denoted using name, TMM values – denoted using name_TMM, and interaction states – denoted using name_int). Dots are patient samples, and the colour of a dot represents the feature value (e.g., difference in BA of the patient compared to the reference value) of that sample for the corresponding feature.

Encoded-SNPs, however, this time incorrect predictions were distributed across the classes rather than both false negatives. Furthermore, as additional validation, we noted that the combined feature set yielded the highest test performance compared to all other tested feature sets when we used a single fixed classifier (Random Forest) (Table S1).

Figure 3e shows the SHAP defined top 15 most predictive features for our best hyper-tuned model generated with AutoKeras, for Integrated-FM-Features, and the relative feature importance's for classification into class 0 (low inflammation, better drug response). We see all three of the feature sets that were concatenated to make Integrated-FM-Features represented in the top 15 features, two features are interaction states, four are TMM values and the remaining nine are patient specific BAs. This highlights the value of our patient-specific BAs and interaction terms in addition to the raw transcriptomic TMM values. The topmost predictive feature (a BA for OR4K15) is the same protein that was identified in Figure S1e where the model was generated using patient specific BAs alone. In fact, similarly to Figure S1e, we observed that that ten of the top fifteen most influential features for predicting drug response were olfactory receptors, with seven distinct genes/proteins represented (since OR5M3 had multiple distinct isoforms represented and OR2H2 is represented as both an interaction term and a TMM).

The second most predictive feature was another olfactory receptor, OR2H2. SNPs within the *OR2H2* gene have been linked to celiac disease in some populations[24], suggesting a possible role in gut and immune function. Notably the second most predictive feature was the interaction term derived from OR2H2 i.e., TMM value multiplied by BA. For OR2H2's BA calculation, 28 patients have one or more of three SNPs (chr6:29,588,032 C/T, chr6:29,588,087 C/T, chr6:29,588,764 C/T) in its associated gene that change the protein sequence (Leu30Phe, Ala48Val, Leu274Phe). 5 patients have no SNPs (and the baseline BA of zero), a further 22 patients have only

one SNP (chr6:29,588,032 that lowers BA to a negative value), 24 patients have two SNPs (chr6:29,588,032 and 29,588,087 that further lower BA), and the remaining patient has all three of the SNPs in this gene which results in the lowest BA recorded. The SNPs drive a lower BA (progressively as SNP number increases) and as such, all the BAs for OR2H2 were negative values or else zero i.e., baseline if no SNPs were seen in the patient. If we consider that all TMM values are positive numbers or zero, all resultant interaction terms for OR2H2 are either negative or zero numbers. As such, in Fig. 3e, a lower (more negative) feature value for the interaction term drives prediction of class 0 (lower inflammation, better response) and indicates a likely higher TMM value and/or a lower – more disrupted BA. We can also see this reflected in the eighth most predictive feature in Fig. 3e (*OR2H2* gene TMM value) where a mid-to-high TMM value directly drives the prediction of class 0. In essence, for prediction of class 0 for OR2H2, the interaction terms highlight the presence of both a change in BA (with increasing SNP presence) and a high associated gene expression level both driving lower inflammation or better drug response.

The third most predictive feature related to the gene *GABRQ*, did not appear previously in Figure S1e in the model trained on BAs alone. The associated protein GABRQ (gamma-aminobutyric acid receptor subunit theta) is a subunit of the GABA-A receptors, ionotrophic channels expressed in both the enteric and central nervous systems, where they bind primary antagonist GABA (gamma-aminobutyric acid)[25,26]. Corticosteroids like prednisolone are not direct ligands of GABA-A receptors[27], however several studies have documented that certain corticosteroids exhibit indirect regulatory effects on GABA-A, likely by influencing the gut microbiota-brain axis[28–30] or influencing production of other GABA receptor modulators[31,32]. In the gut GABA receptor activation has also been shown to inhibit pro-inflammatory cytokine secretion and increase anti-inflammatory cytokine production[33,34], highlighting it's role in regulating intestinal inflammation. General evidence suggests that increased GABAergic signalling (potentially linked to higher receptor function or gene expression) is associated with reduced inflammation, while corticosteroids tend to inhibit this system. Interestingly in Fig. 3e, lower *GABRQ* expression drives prediction of class 0 (better response to prednisolone, lower inflammation). It is plausible that variations in a GABA-A receptor subunit, could influence an individual's overall GABAergic tone, indirectly affecting the balance of immune and inflammatory responses, potentially altering the efficacy of treatments like prednisolone.

This gene *GABRQ* is therefore an interesting candidate and similarly to *OR2H2* appears twice in the most predictive features as a TMM value and an interaction state, although here the TMM value is the more predictive of the two feature sets. When we look at the interaction state a higher interaction drives a class 0 prediction. Examination of the BAs for the GABRQ protein driving the BA calculation revealed the presence of one SNP (chrX:152,652,814, A/T) in 19 patients, which resulted in a change in protein sequence (Ile478Phe) and decreased the protein-drug BA compared to the other patients exhibiting no SNPs that had a baseline BA of zero. The SNPs drive a lower BA and as such, similarly to for OR2H2, all the BAs for GABRQ are negative values or else zero, as are all resultant interaction terms. As such, in Fig. 3e, a higher (typically zero) feature value for the interaction term drives prediction of class 0 (lower inflammation, better response) and this zero interaction state value results from; a baseline BA only (unmutated GABRQ) in 45% of patients, a TMM value of zero in 24% of patients and from both a baseline BA and unexpressed gene in the remaining 31% of patients. Lower expression of *GABRQ* driving over half of the class 0 predictions here supports the pattern seen for the TMM value feature directly.

Using model context protocol (MCP) and AI agents to make BA calculations accessible

We note that even when a user is using a FM for inference (rather than fine-tuning or pre-training), the construction and execution of a FM workflow can be a time consuming and interdisciplinary task, confining usage to specialised computational experts. For example, workflows move from processing raw omics or molecular data, through FM inference, to interpretation of the meaning and biological validity of the results. AI Agents are Large Language Model (LLM)-driven systems which autonomously plan, reason and dynamically call tools/functions. They provide opportunities to accelerate biomedical discovery for non-programmers or non-AI experts, since they can remove coding and computational pre-requisites, allowing natural language prompting to execute complex tasks. As an example of this, we use an AI Agent to make a FM inference call and to calculate the patient-specific BAs that we derive and use in this study. To achieve this, we combine a popular community AI agent (Claude, available at https://claude.ai/) with a MAMMAL FM MCP server (accessible at: https://github.com/BiomedSciAI/biomed-multi-alignment/tree/main/mammal_mcp) that we developed to make our FM drug-target BA inference task useable by the AI Agent. Furthermore, we provide step-by-step instructions and test datasets to ensure users can harness this workflow and capability for themselves (see Methods).

Figure 4 shows the results of such an agent-driven analysis to compare prednisolone BA for a sample set of three mutated proteins from an imagined patient to the unmutated reference set of those same proteins. For exemplar testing we used the two known targets of prednisolone (NR3C1 and NR3C2) plus our most predictive protein from Figure S1e and 3e (OR4K15). Our imagined patient is given all the observed mutations in NR3C1/2 across our cohort (see Methods). Figure 4 shows that even this highest mutational load in the associated gene for these proteins had little impact on BA, potentially why these proteins (as SNPs or BAs) did not appear as predictive of drug response in any of our models. In contrast, the OR4K15 protein was mutated at three missense SNP positions that were observed in the patients (see Methods) and this had much greater impact on BA, supporting the value of this feature for drug response prediction. Our agent, Claude in this instance, correctly calls MAMMAL, performs inference, calculates delta changes in BA between patients and the reference and clearly summarises the findings with minimal prompting. Interestingly it also detects the corticosteroid-like nature of the input compound and the functions of our input proteins e.g., receptor types, without user prompting. We propose that a user could switch to another drug of interest or change input proteins with relative ease using this approach.

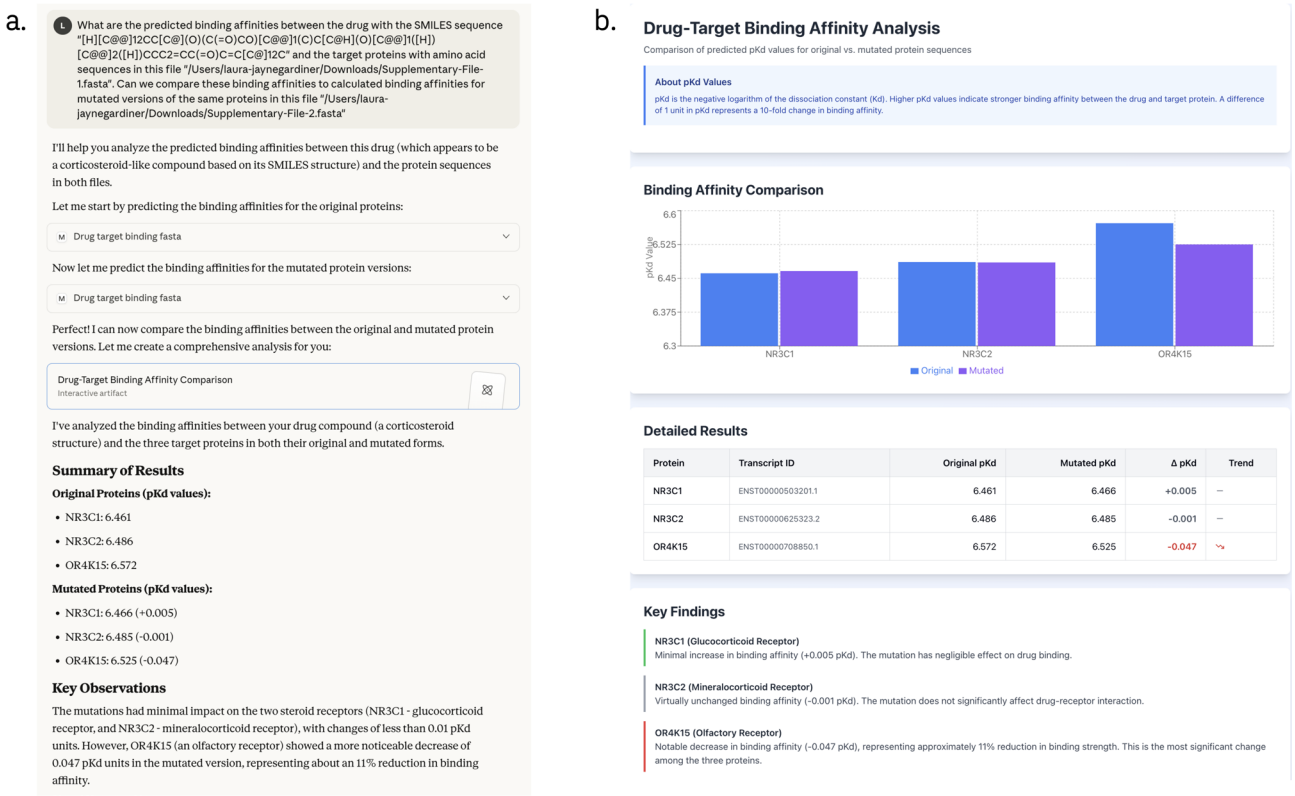

Fig. 4. Output from AI agent (Claude) making patient specific BA calculations. (**a**) Primary output from Claude after input of shown user prompt (also defined in Methods) to enable comparison of mutated patient proteins to unmutated reference (baseline) proteins for the calculation of BA with Prednisolone using FM inference. (**b**) Claude detailed tabular and graphical output summarising results further, accessed after clicking on the "Drug-Target Binding Affinity Comparison – Interactive artifact" button in (**a**).

Conclusion

This work used a FM that had been fine tuned to perform drug-target BA inference to enable feature selection and to engineer multi-omics features that allowed more accurate prediction of drug response in IBD patients. In doing so we were able to build AI/ML driven workflows that suggested novel associations between specific SNPs, genes and proteins, and personalized drug response to prednisolone. Although we showcase a single FM inference task in this study, our approach is not limited to MAMMAL, or to this task, in fact any FM providing a relevant measure that could be used as a feature to predict drug response is appropriate e.g., a different FM tuned for drug-target BA, FMs tuned for protein–protein interaction or T-cell receptor binding from sequence/SMILES inputs.

While a high volume of functional and genomics data was generated for this study, and the sample number was relatively high for an ex vivo study of human fresh tissues, we note that the total number of patients is low for an association study. Therefore, we prioritised considerations around the application of ML to small sample numbers and the validity of the biological findings are tentative. Further work is required to confirm and validate these associations; however, our aim was to extract new features that might be informative for learning drug-related targets from genomics and transcriptomics data, and as such, the cohort acted as a consistent and practical baseline for comparative analyses. Following extensive validation, such pharmacogenomic approaches could enable patient stratification for personalised drug treatment. It is likely that most of our patient samples used in our organoculture assay have already been exposed to prednisolone previously, which could impact their responsiveness in vitro, which represents a further limitation of this study.

This work effectively demonstrates the potential of using FM inference to enhance personalized medicine discoveries. We demonstrate the downstream application of a pre-trained FM to gain advantage for a small multi-omics cohort, leveraging the FM's generalised learning over large populations. To the best of our knowledge, comparison to related works suggests that this is a novel approach for biomedical FM-enabled feature generation for drug response prediction. Furthermore, using a protein model for a primarily genomics and transcriptomics analysis, we show the possibility to extend FM application beyond its intended trained modalities. Finally, by combining AI agents (where we use Claude as an example) with a custom built MCP server (hosting our fine-tuned FM), we make complex FM inference aspects of our workflow accessible to non-experts via natural language prompts.

Although this study focuses on paired bulk transcriptomic and genomic data, our workflow is inherently modality-agnostic since the FM-driven steps—drug–target BA inference and protein-level feature engineering—

operate independently of additional omics layers. As a result, the same FM-prioritised protein features and interaction-state representations can be directly mapped onto single cell RNA-seq, spatial transcriptomics, or other modalities. For single-cell or spatial data, gene-expression values for the FM-ranked proteins would simply replace bulk values, enabling interaction-state features to be computed per cell, per patient-level aggregation or per spatial location. Furthermore, our workflow allows incorporation of additional modalities such as ATAC-seq, proteomics, either as separate feature channels or through analogous interaction-state representations. We therefore view the workflow as broadly extensible and anticipate that applying it to other multi-omics datasets may further enhance biological interpretability e.g., adding cellular or spatial resolution.

Methods

Exome sequencing and SNP analysis

Whole exome sequencing and RNA-seq plus downstream bioinformatics (alignment and SNP calling) were performed for the 51 patients as described in our previous work, except this time for 51 patients i.e., adding 27 to the 25 patients reported previously[18]. Genomic data was provided as a BAM file aligned to genome reference GRCh37. Genotypes called with Torrent Variant Caller were provided as per sample VCF files. Single nucleotide polymorphisms (SNPs) from the VCF files were merged into a multi-sample VCF and VCF files were then filtered to remove low quality SNPs including those with a depth less than 30. During SNP post-processing, we translated SNPs from the GRCh37 reference genome coordinates to the GRCh38 reference genome coordinates using the Ensembl Assembly Converter (https://grch37.ensembl.org/Homo_sapiens/Tools/AssemblyConverter). In total, 1,948,348 unique SNP positions were identified across the patients.

The software snpEff (v4.3t)[35] was used to annotate the SNPs within the VCF files to enable downstream selection of missense SNPs that would impact protein amino acid sequence. To enable downstream AI/ML directly from the SNPs, we labelled SNPs as: Homozygous (LABEL = 2) if their alternate allele frequency was > 0.8; Heterozygous (LABEL = 1) if their alternate allele frequency was < = 0.8; Homozygous REF (LABEL = 0) if no SNP was recorded for that patient at that position. Next, we filtered indels and SNPs where > 50 (95%) of the patients showed the same allele (leaving 169,951 SNPs).

RNA-seq bioinformatic analysis

Raw paired-end reads were obtained in FASTQ format (un-stranded) for the 51 patients. These reads were ran through the nf-core/rnaseq workflow[36] (v3.14.0). This pipeline quantifies RNA-sequenced reads relative to genes/transcripts in the genome and normalizes the resulting data. The workflow was run for all of the 51 samples implementing the default method (QC, trimming, STAR and Salmon) and using the Human GRCh38 reference genome. We identified a relatively high degree (47% on average of reads per sample) of non-optical or PCR duplicates across the samples. However, since the remaining number of unique (non-duplicate) reads per sample falls at an average of 53M, with all samples having above 20M reads, the sequencing appears to be perfectly adequate for analysis requirements. **T**he vast majority of reads (average 87% across the samples) were aligned uniquely to the reference genome. The gene count matrix that is given as output from this bioinformatics workflow was used as input into EdgeR to generate normalized values according to the weighted trimmed mean of the log expression ratios i.e., trimmed mean of M values (TMM values). Retaining only those genes that are represented in at least 10% of the samples i.e., approximately 5 or more patients, resulted in a matrix of 51 patients and 23,139 genes.

Feature selection

In the case of finding a specified number of best features using a classic or standard approach, we used the *SelectKBest* method (*f classif* metric) as implemented by scikit learn[37]. *SelectKBest* works by taking the scoring metric provided to it and selecting the K best scoring features. If using *f classif*, this is the ANOVA *F*-value of the column in relation to the targets. In the case of our second approach to feature selection deriving feature rankings directly using an inference task from the FM MAMMAL that has been fine tuned to predict drug-target binding or pKd. Specifically, we infer BAs for each unique reference protein amino acid sequence in the GRCh38 reference genome against the SMILES representation of prednisolone. To derive a representative set of unique human proteins in the GRCh38 reference, overlapping the requirements of MAMMAL for sequence lengths under 1250 bp we focused on those proteins over 200aa in length and less than 1240aa (37,517 of which 11,417 unique genes were represented so we selected the primary protein isoform labeled "0.1" for each gene). Each of the 11,417 proteins was run in turn through the MAMMAL drug-target binding affinity model for inference to assess its BA to our target drug prednisolone. For reference, the SMILES representation that we used for our target drug of interest prednisolone is as follows: **"**[H][C@@]12CC[C@](O)(C(=O)CO)[C@@]1(C)C[C@H](O)[C@@]1([H])[C@@]2([H])CCC2=CC(=O)C=C[C@]12C" that was taken directly from the website Drugbank (https://go.drugbank.com/drugs/DB00860).

Proteins were then ranked according to the extent of their BA with prednisolone as calculated using FM inference and we selected the top 5% of most tightly bound reference proteins (rounded up to 600). The 5% threshold was selected based on a combination of established statistical convention and empirical validation. Statistical Convention**:** The 5% cutoff is a widely accepted significance threshold in biological and biomedical research, representing a balance between stringency and sensitivity. This convention facilitates comparison and maintains consistency with standard practices in drug-target interaction studies. Empirical Validation: The appropriateness of the 5% threshold was confirmed by the inclusion of known targets of the drug within this top percentile. This concordance between the selected threshold and experimentally validated targets provides strong empirical support that the cutoff successfully enriches for biologically relevant interactions while minimizing false positives. Distribution Characteristics: Examination of the BA distribution (Fig. 2a) reveals a natural inflection point around the 95th percentile (a BA of approx. 6.5), where high-affinity interactions

become distinct from the broader population. The top 5% also captures proteins with BAs substantially above the population mean, suggesting meaningful biological interactions. Alternative thresholds (e.g., 6% or 10%) would increase sensitivity but at the cost of reduced specificity, potentially introducing lower-confidence interactions that dilute the dataset. Conversely, a more stringent threshold (e.g., 2–3%) risks excluding true positive interactions with slightly lower but still biologically relevant binding affinities.

We performed enrichment analysis for these 600 human reference proteins using WebGestalt 2024[38] against a background or reference including all protein coding genes in the genome. We performed over-representation analysis including multiple test adjustment and the top categories were selected after FDR ranking. Missense SNPs were identified per patient from the genomics data and were used to mutate the 600 proteins (python package Bio.seq.MutableSeq), resulting in a per-patient mutated protein set. BAs of mutated proteins with prednisolone were then predicted using the fine-tuned MAMMAL model. A final patient-specific BA was then calculated by subtracting the reference BA from the patient mutated protein BA.

Machine learning (ML)

To perform ML we used the open source tool AutoXAI4Omics[22]. We used this in classification mode, comparing a range of classifiers (Random Forest, AdaBoost, KNN, AutoXGBoost, AutoLightGBM, AutoKeras and AutoSKlearn), using a seed of "80", a train/test split of 80/20 and a random search for hyperparameter tuning (rather than default grid search). AutoXAI4Omics uses k-fold cross-validation during model training (specifically, stratified k-fold cross-validation for classification tasks with 3-folds used for this dataset). Our feature sets and feature selection methods are detailed in the main text and methods above. Our label to predict relates to TNFα levels (pg/mL) that were determined for each patient sample in the organoculture assay described in our previous work[18], these levels were normalised against the patient sample matched vehicle control group. The individual drug treated biopsy results were calculated as a percentage of the mean vehicle control group results before both drug treated duplicate biopsy results were then meaned to provide a single result for each drug treatment per donor sample. **T**he TNFα ranges were as follows for 1 µM Prednisolone treatment 36.8–293.8, and we derived classes from this data equating to high (class 1) and low TNFa levels (class 0) for patients with a TNFα level greater than or equal to 82 and less than or equal to 79 respectively. This amounted to a relatively balanced set of 29 patients in class 1 and 22 patients in class 0.

Using agents to make binding affinity calculations accessible

We use an AI Agent to make a series of FM inference calls and to calculate resulting patient-specific BAs (with Prednisolone) using the same method that we derive and use in this study. To achieve this, we combine the ClaudeAI agent (Claude, available at https://claude.ai/, other MCP compatible AI agent systems would work like ChatGPT and Langflow.) with our developed MAMMAL FM MCP server (at: https://github.com/BiomedSciAI/biomed-multi-alignment/tree/main/mammal_mcp) to make our FM drug-target BA inference task accessible to the AI Agent. As an exemplar task we take three unmutated proteins from our identified set of 600 that had the highest BAs with Prednisolone (Supplementary-File-1-fasta.txt). These three unmutated proteins include the two known targets of Prednisolone (NR3C1 and NR3C2) plus our most predictive protein from Figs. 3e and 4e (OR4K15). For comparison, we have the same three proteins post-mutation (Supplementary-File-2-fasta.txt). In Supplementary-File-2, NR3C1 is mutated at three positions of the missense SNPs that were observed in this gene across our 51 patients (chr5:143,399,752, T/C, Asn363Ser; chr5:143,400,590, G/C, Leu84Val; chr5:143,400,772, C/T, Arg23Lys); NR3C2 is mutated at the one missense SNP position that was observed in this gene across our 51 patients (chr4:148,436,323, C/T, Val180Ile); OR4K15 is mutated at three missense SNP positions that were observed (chr14:19,975,784, A/G, Asn89Ser; chr14:19,975,867, T/G, Ser117Ala; chr14:19,976,282, C/A, Ala255Glu). To replicate the process that we followed to ingest these sequences and calculate BA, a user must:

1. Download Claude desktop and install (https://claude.com/download)
2. Go to the "biomed-multi-alignment" Github repo and follow installation instructions. Or "git clone https://github.com/BiomedSciAI/biomed-multi-alignment.git"
3. Change directory into your newly created "biomed-multi-alignment" folder. Then change directory again into the sub-folder "mammal_mcp".
4. On Github navigate to the "mammal_mcp" folder and follow the "Getting started" instructions to run the MCP server and then follow the instructions to allow "Integration into Claude Desktop" (at: https://github.com/BiomedSciAI/biomed-multi-alignment/tree/main/mammal_mcp)
5. Download the test fasta files Supplementary-File-1-fasta.txt (unmutated proteins) and Supplementary-File-2-fasta.txt (mutated proteins). Rename these as Supplementary-File-1.fasta and Supplementary-File-2.fasta respectively.
6. Open Claude desktop and ask the following question:

   What are the predicted binding affinities between the drug with SMILES sequence "[H][C@@]12CC[C@](O)(C(=O)CO)[C@@]1(C)C[C@H](O)[C@@]1([H])[C@@]2([H])CCC2=CC(=O)C=C[C@]12C" and the target proteins with amino acid sequences in this file "/Full-path-to-file/Supplementary-File-1.fasta". Can we compare these binding affinities to calculated binding affinities for mutated versions of the same proteins in this file ""/Full-path-to-file/Supplementary-File-2.fasta"

Note that this Claude workflow uses the MAMMAL model "ibm/biomed.omics.bl.sm.ma-ted-458m.dti_bindingdb_pkd" rather than the specific version "ibm/biomed.omics.bl.sm.ma-ted-458m.dti_bindingdb_pkd_peer" that was used in this study by default – this can be adjusted to emulate identical BAs to our study.

Data availability

The dataset that was generated during the current study and used to train our best ML model is available in Supplementary File 3 (calculated Binding affinities, TMM values and interaction states, 1230 features for the 51 patients). The sequences used with our example AI agent workflow are also available in Supplementary Files 1 and 2 (fasta sequences available as txt files). Where consent has been provided, the raw RNA datasets analysed during the current study are available in the NCBI-SRA repository, under project ID PRJEB43220. Further data and metadata are available through a controlled access route to maintain the requirements of ethical compliance. Applications for access can made through the manuscript authors affiliated with REPROCELL-Europe Ltd.

Code availability

The MCP server code to access our FM inference task is open-source at: https://github.com/BiomedSciAI/biomed-multi-alignment/tree/main/mammal_mcp. The code to carry out our bioinformatics processing (genomics, SNP annotation, mutation of proteins, RNA-seq analysis) is all based on open source tools and packages that are detailed in the methods.

Received: 9 December 2025; Accepted: 11 March 2026
Published online: 17 March 2026

References

1. Vaswani, A. et al. Attention Is All You Need. Preprint at https://doi.org/10.48550/ARXIV.1706.03762 (2017).
2. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Preprint at https://doi.org/10.48550/ARXIV.1810.04805 (2018).
3. Ji, Y., Zhou, Z., Liu, H. & Davuluri, R. V. DNABERT: Pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. *Bioinformatics* **37**, 2112–2120 (2021).
4. Dandala, B. et al. BMFM-RNA: an open framework for building and evaluating transcriptomic foundation models. Preprint at https://doi.org/10.48550/ARXIV.2506.14861 (2025).
5. Yang, F. et al. scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data. *Nat. Mach. Intell.* **4**, 852–866 (2022).
6. Cui, H. et al. scGPT: Toward building a foundation model for single-cell multi-omics using generative AI. *Nat. Methods* **21**, 1470–1480 (2024).
7. Theodoris, C. V. et al. Transfer learning enables predictions in network biology. *Nature* **618**, 616–624 (2023).
8. Abramson, J. et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500 (2024).
9. Shoshan, Y. et al. MAMMAL -- Molecular aligned multi-modal architecture and language. Preprint at https://doi.org/10.48550/ARXIV.2410.22367 (2024).
10. Stankey, C. T. et al. A disease-associated gene desert directs macrophage inflammation through ETS2. *Nature* **630**, 447–456 (2024).
11. King, E. A., Davis, J. W. & Degner, J. F. Are drug targets with genetic support twice as likely to be approved? Revised estimates of the impact of genetic support for drug mechanisms on the probability of drug approval. *PLoS Genet.* **15**, e1008489 (2019).
12. De Lange, K. M. et al. Genome-wide association study implicates immune activation of multiple integrin genes in Inflammatory Bowel Disease. *Nat. Genet.* **49**, 256–261 (2017).
13. Kumar, M., Garand, M. & Al Khodor, S. Integrating omics for a better understanding of Inflammatory Bowel Disease: A step towards personalized medicine. *J. Transl. Med.* **17**, 419 (2019).
14. Bradley, J. TNF-mediated inflammatory disease. *J. Pathol.* **214**, 149–160 (2008).
15. Plevy, S. E. et al. A role for TNF-alpha and mucosal T helper-1 cytokines in the pathogenesis of Crohn's disease. *J. Immunol.* **159**, 6276–6282 (1997).
16. Olsen, T. et al. Tissue levels of tumor necrosis factor-alpha correlates with grade of inflammation in untreated ulcerative colitis. *Scand. J. Gastroenterol.* **42**, 1312–1320 (2007).
17. Mandrekar, S. J. & Sargent, D. J. All-comers versus enrichment design strategy in Phase II trials. *J. Thorac. Oncol.* **6**, 658–660 (2011).
18. Gardiner, L.-J. et al. Combining explainable machine learning, demographic and multi-omic data to inform precision medicine strategies for inflammatory bowel disease. *PLoS ONE* **17**, e0263248 (2022).
19. Flegel, C., Manteniotis, S., Osthold, S., Hatt, H. & Gisselmann, G. Expression profile of ectopic olfactory receptors determined by deep sequencing. *PLoS ONE* **8**, e55368 (2013).
20. Evans, P. D., Bayliss, A. & Reale, V. GPCR-mediated rapid, non-genomic actions of steroids: Comparisons between DmDopEcR and GPER1 (GPR30). *Gen. Comp. Endocrinol.* **195**, 157–163 (2014).
21. Kaplan, I., Blakely, B. T., Pavlath, G. K., Travis, M. & Blau, H. M. Steroids induce acetylcholine receptors on cultured human muscle: Implications for myasthenia gravis. *Proc. Natl. Acad. Sci. U.S.A.* **87**, 8100–8104 (1990).
22. Strudwick, J. et al. AutoXAI4Omics: An automated explainable AI tool for omics and tabular data. *Brief. Bioinform.* **26**, bbae593 (2024).
23. Sollai, G. et al. Olfactory function in patients with inflammatory bowel disease (IBD) is associated with their body mass index and polymorphism in the odor binding-protein (OBPIIa) gene. *Nutrients* **13**, 703 (2021).
24. Banerjee, P. et al. Association study identified biologically relevant receptor genes with synergistic functions in celiac disease. *Sci. Rep.* **9**, 13811 (2019).
25. Auteri, M., Zizzo, M. G. & Serio, R. GABA and GABA receptors in the gastrointestinal tract: From motility to inflammation. *Pharmacol. Res.* **93**, 11–21 (2015).
26. Sinkkonen, S. T., Hanna, M. C., Kirkness, E. F. & Korpi, E. R. $GABA_A$ receptor ε and θ subunits display unusual structural variation between species and are enriched in the rat Locus Ceruleus. *J. Neurosci.* **20**, 3588–3595 (2000).
27. Phulera, S. et al. Cryo-EM structure of the benzodiazepine-sensitive α1β1γ2S tri-heteromeric GABAA receptor in complex with GABA. *Elife* **7**, e39383 (2018).
28. So, S. Y. & Savidge, T. C. Gut feelings: The microbiota-gut-brain axis on steroids. *Am. J. Physiol. Gastrointest. Liver Physiol.* **322**, G1–G20 (2022).
29. Tetel, M. J., De Vries, G. J., Melcangi, R. C., Panzica, G. & O'Mahony, S. M. Steroids, stress and the gut microbiome-brain axis. *J. Neuroendocrinol.* **30**, e12548 (2018).
30. McCurry, M. D. et al. Gut bacteria convert glucocorticoids into progestins in the presence of hydrogen gas. *Cell* **187**, 2952-2968.e13 (2024).
31. De Lima, A. M. D. L. et al. Effect of prednisolone in a kindling model of epileptic seizures in rats on cytokine and intestinal microbiota diversity. *Epilepsy Behav.* **155**, 109800 (2024).

32. Di, S., Maxson, M. M., Franco, A. & Tasker, J. G. Glucocorticoids regulate glutamate and GABA synapse-specific retrograde transmission via divergent nongenomic signaling pathways. *J. Neurosci.* **29**, 393–401 (2009).
33. Deng, Z. et al. Activation of GABA receptor attenuates intestinal inflammation by modulating enteric glial cells function through inhibiting NF-κB pathway. *Life Sci.* **329**, 121984 (2023).
34. Ma, X. et al. Activation of GABAA receptors in colon epithelium exacerbates acute colitis. *Front. Immunol.* **9**, 987 (2018).
35. Cingolani, P. et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly (Austin)* **6**, 80–92 (2012).
36. Patel, H. et al. nf-core/rnaseq: nf-core/rnaseq v3.21.0 - Mercury Macaw. Zenodo https://doi.org/10.5281/ZENODO.1400710 (2025).
37. Pedregosa, F. et al. Scikit-learn machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
38. Elizarraras, J. M. et al. WebGestalt 2024: Faster gene set analysis and new support for metabolomics and multi-omics. *Nucleic Acids Res.* **52**, W415–W421 (2024).

Acknowledgements
This work was supported by the Hartree National Centre for Digital Innovation (HNCDI), a collaboration between STFC and IBM.

Author contributions
All authors contributed code, bioinformatics support, and/or study conceptualization. L.J.G. performed primary conceptualisation, analyses and manuscript writing. J.K. performed MCP server development and interaction state determination. A.E. performed MCP server development. S.C. performed transcriptomics bioinformatics. L.J.G, J.K., A.E., G.M., D.B., K.B., performed manuscript review and editing.

Funding
This work was supported by the Hartree National Centre for Digital Innovation, a collaboration between the Science and Technologies facilities Council (STFC) and IBM (LJG, J.K, A.E, S.C) that is funded by the UK Research and Innovation (UKRI) funding agency. The functional pharmacology experiments and whole exome sequencing for this study were part-funded via a project grant from Precision Medicine Scotland Innovation Centre (PMS-IC), provided to REPROCELL Europe Ltd (K.B, G.M, D.B). PMS-IC is funded by the Scottish Funding Council and Scottish Enterprise. The funding organisations did not play an additional role in the study design, data collection and analysis, or preparation of the manuscript and only provided financial support in the form of authors' salaries and research materials.

Declarations

Competing interests
The authors declare no competing interests.

Ethics approval and consent to participate
The research was conducted with the approval of the West of Scotland Research Ethics Committee (approvals 12/ws/0069, 17/WS/0049 and 22/WS/0007) and the Advarra Institutional Review Board (Protocol ID: Pro00005300). All procedures performed in this study involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards. Informed consent and written consent was obtained from all individual participants involved in the study.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-026-44366-y.

Correspondence and requests for materials should be addressed to L.-J.G.