International Journal of Pharmaceutics 634 (2023) 122643

Contents lists available at ScienceDirect

International Journal of Pharmaceutics

journal homepage: www.elsevier.com/locate/ijpharm

Advancing oral delivery of biologics: Machine learning predicts peptide stability in the gastrointestinal tract

Fanjin Wang [a], Nannapat Sangfuang [b], Laura E. McCoubrey [b], Vipul Yadav [a], Moe Elbadawi [b], Mine Orlu [b], Simon Gaisford [b], Abdul W. Basit [b,*]

[a] Intract Pharma Ltd. London Bioscience Innovation Centre, 2 Royal College St, London NW1 0NH, UK
[b] UCL School of Pharmacy, 29-39 Brunswick Square, London WC1N 1AX, UK

ARTICLE INFO

Keywords:
Artificial intelligence
Biopharmaceuticals
Degradation of peptides and proteins
Formulation and delivery of macromolecules
Gastrointestinal absorption and bioavailability
Drug development and industry 4.0

ABSTRACT

The oral delivery of peptide therapeutics could facilitate precision treatment of numerous gastrointestinal (GI) and systemic diseases with simple administration for patients. However, the vast majority of licensed peptide drugs are currently administered parenterally due to prohibitive peptide instability in the GI tract. As such, the development of GI-stable peptides is receiving considerable investment. This study provides researchers with the first tool to predict the GI stability of peptide therapeutics based solely on the amino acid sequence. Both unsupervised and supervised machine learning techniques were trained on literature-extracted data describing peptide stability in simulated gastric and small intestinal fluid (SGF and SIF). Based on 109 peptide incubations, classification models for SGF and SIF were developed. The best models utilized k-Nearest Neighbor (for SGF) and XGBoost (for SIF) algorithms, with accuracies of 75.1% (SGF) and 69.3% (SIF), and f1 scores of 84.5% (SGF) and 73.4% (SIF) under 5-fold cross-validation. Feature importance analysis demonstrated that peptides' lipophilicity, rigidity, and size were key determinants of stability. These models are now available to those working on the development of oral peptide therapeutics.

1. Introduction

Peptide-based drugs represent a significant class of biological treatments, with market-leading successes including liraglutide (Victoza®), goserelin (Zoladex®), and leuprolide (Lupron®). Due to their comparative structural complexity, peptide-based drugs typically facilitate enhanced target specificity compared to conventional small molecule drugs, affording higher therapeutic success and reduced off-target effects (Camela et al., 2021; Lasa et al., 2022). Despite their many advantages, peptide-based drugs are often unstable in the gastrointestinal (GI) tract, requiring over 90 % of marketed peptide therapeutics to be administered parenterally (Kremsmayr et al., 2022). Parenteral administration is a key driver of the high costs and reduced accessibility associated with biopharmaceutical treatments, as healthcare professionals must typically be present for the administration of each dose (Makurvet, 2021). Further, injections are less acceptable to patients than oral formulations, with injection frequency particularly associated with lower health-related quality of life (Boye et al., 2011). For these reasons, the development of orally-administered peptide therapeutics presents a key opportunity to improve current treatment strategies for numerous diseases (Abramson et al., 2022; Zhang et al., 2021).

A considerable challenge facing oral peptide delivery is low bioavailability due to poor peptide stability and/or permeability in the GI tract. Peptide physicochemical characteristics determine susceptibility to GI degradation and permeability across the epithelium (Klepach et al., 2022; Lau and Dunn, 2018). By optimizing GI peptide stability, peptides are available for local therapeutic action and are more likely to reach the epithelium intact for systemic access. There are two main barriers facing the GI stability of peptides: gastric acid and intestinal enzymes (Wang et al., 2015b). Due to their selectivity, it is imperative that peptide drugs maintain a conformation that allows them to interact with their physiological target. The high concentration of protons in gastric fluid can denature therapeutic peptides by destabilizing their secondary and tertiary structures through the disruption of hydrogen

Abbreviations: CV, cross-validation; DT, Decision tree; GI, gastrointestinal; KNN, K nearest neighbor; ML, machine learning; PCA, principal component analysis; RF, random forest; RFE, recursive feature elimination; SGF, simulated gastric fluid; SIF, simulated intestinal fluid; SMILES, simplified molecular-input line-entry system; SVM, support vector machine; TPSA, total polar surface area; USP, United States Pharmacopeia.

* Corresponding author.
E-mail address: a.basit@ucl.ac.uk (A.W. Basit).

https://doi.org/10.1016/j.ijpharm.2023.122643
Received 5 December 2022; Received in revised form 18 January 2023; Accepted 20 January 2023
Available online 25 January 2023

and ionic bonds (Wicke et al., 2021). Gastric acid also activates pepsin, an enzyme that can cleave peptides into inactive fragments (Kremsmayr et al., 2022). Moreover, both human and microbial enzymes in the small and large intestine can inactivate peptides. For example, pancreatic secretions entering the small intestine contain peptidases with broad substrate specificity, including enzymes that can cleave peptides at aromatic, charged, and neutral residues (Ahmed et al., 2022; Whitcomb and Lowe, 2007). The intestinal microbiota also produces enzymes with wide functionality that can chemically inactivate small molecule and peptide drugs alike (McCoubrey et al., 2021). Whilst microbiota do colonise the small intestine, the colon houses the highest density of microorganisms of the entire body, thus biopharmaceuticals targeted to the colon are at particular risk of microbial inactivation (McCoubrey et al., 2022a; Yadav et al., 2016).

At present there is no validated method for predicting the GI stability of peptides intended for oral delivery (Drucker, 2020). Whilst prior research has revealed that engineering peptide backbones with unnatural amino acids, D-amino acids, cyclisation, polymer conjugation, and increasing lipophilicity can protect peptides from enzymatic inactivation in intestinal fluids, the extent that these modifications can improve stability has not been broadly quantified (Drucker, 2020; Elfgen et al., 2019; Zizzari et al., 2021). One reason that a predictive technology does not yet exist is the extensive molecular space available for peptide design (Narayanan et al., 2021). Peptides are proteins containing 2 – 50 amino acids, conferring considerably greater structural complexity and possible configurations than small molecule drugs (Drucker, 2020; Forbes and Krishnamurthy, 2022). As such, the comprehension of how such a large chemical space maps to interactions with the multifarious components of GI fluids has eluded human assessment; traditional tools used for small molecule drugs, such as Lipinski's Rule-of-5, are less reliable for peptides (Brayden et al., 2020; Lohman et al., 2019; Nielsen et al., 2017). Here artificial intelligence (AI) can be utilized to identify important trends within dense datasets and output predictions for untested peptides' stabilities (Narayanan et al., 2021). Machine learning (ML), a form of AI, has been widely harnessed in recent years to predict protein structure and functionality, such as protein binding affinity to specified targets (Gao et al., 2020; Jumper et al., 2021). ML has great potential for detecting how slight nuances in peptide structure can impact stability in the GI tract; the advantages of such knowledge include pre-clinical prediction of peptide suitability for oral delivery and design of novel highly stable peptide structures (Chandrasekaran et al., 2018; Gao et al., 2017).

In this study, various ML strategies were developed and compared for their ability to predict the stability of peptide drugs in simulated gastric and small intestinal fluid (SGF and SIF, respectively) using peptide structure as an input. The training dataset was constructed using a strategic literature mining approach and learning performance was benchmarked against a baseline model that reported the arithmetic mode (i.e. the most frequent class) of peptide stability in the training dataset. The findings reveal important structure-stability relationships for the design of novel oral peptide therapeutics and facilitate the prediction of any untested peptide's stability in the human GI tract. The optimized predictive models are available online at: https://github.com/FrankWanger/ML_Peptide.

2. Materials and methods

A schematic representation of this study's pipeline is shown in Fig. 1.

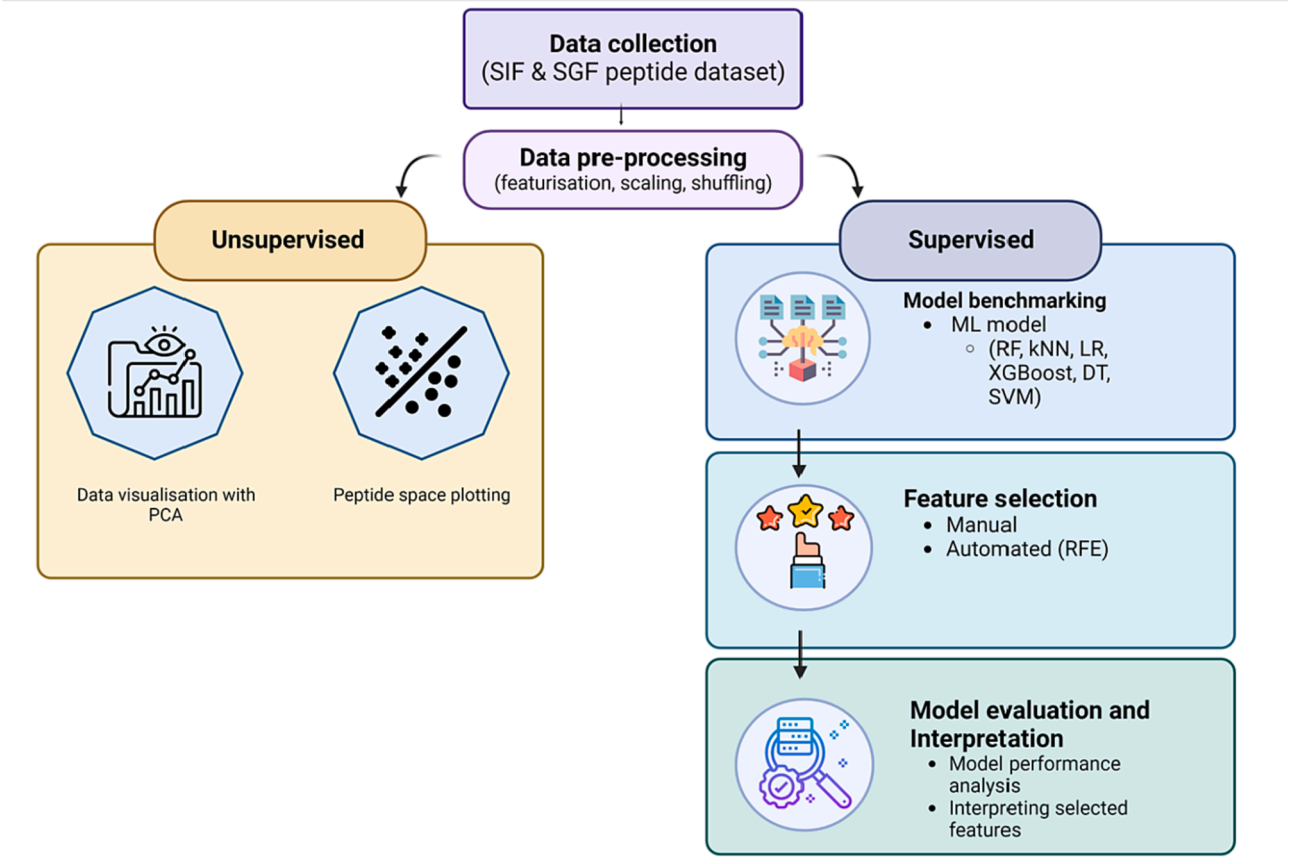

Fig. 1. A schematic representation of the study's pipeline. Unsupervised: unsupervised machine learning to explore relationships between the physicochemical properties of the peptides included in the dataset; Supervised: supervised machine learning to build predictive models that take physicochemical properties as inputs and output predictions on gastrointestinal stability; PCA: principal component analysis; RF: random forest; kNN: k-nearest neighbor; LR: logistic regression; DT: decision tree; SVM: support vector machine; RFE: recursive feature elimination.

Peptide stability data was first identified and extracted from online literature, followed by data featurization and pre-processing. Unsupervised learning, ML that does not require the target class to be labelled (Bell, 2022), was first conducted to visualize the database and to compare extracted peptides with a larger 'peptide space' formed by the U.S. Food and Drug Administration (FDA)-approved therapeutic peptides. Then, supervised learning algorithms, which utilize labelled data to form predictions (Bell, 2022), were benchmarked, and feature selection was conducted to further improve model performances. Finally, model classification performances were analyzed and feature interpretation was conducted to shed light on the knowledge obtained by the ML models.

2.1. Database preparation

2.1.1. Data collection

The peptide stability training dataset was generated by retrieving the stability of peptides in SGF and SIF from the literature, identified via PubMed and Google Scholar. The SGF and SIF used in the literature studies were prepared according to USP guidelines and simulate the physiological pH of GI tract in the fasted state. Results generated from incubations with SGF and SIF were sought (for example, over stability data from animal models) because SGF/SIF stability is commonly assayed in industry, there is considerably more available data describing SGF/SIF peptide stability, and SGF stability has been reported as correlating well with gastric intestinal stability in humans ($R^2 = 0.917$) (Wang et al., 2015b), whereas animal GI physiology can significantly differ to that of humans (Hatton et al., 2015; Kremsmayr et al., 2022). Although there is no published correlation between SIF and human intestinal fluid for peptide stability, SIF is relevant due to its prominence in the pharmaceutical industry for oral drug development, for example in dissolution testing (Bou-Chacra et al., 2017).

Specific search terms and the year of publication were used to find studies examining the *in vitro* stability of peptides. The key searching terms were 'peptides', 'stability', 'peptide drugs', 'SGF', 'SIF', 'chemical stability' and 'GI tract'. The search operator words used for searching were 'AND' and 'OR'. The specific search terms alongside the number of study results are presented in Table 1. The context and quality of each study was investigated by domain specialists to assure its relevancy before addition to the training dataset. The complete data selection and extraction process is presented in Fig. 2.

2.1.2. Data extraction

The majority of peptide stability data in SGF and SIF were directly acquired from publications as raw data. However, some stability data required extraction from figures. Here, the online tool WebPlotDigitizer (version 4.5 developed by PLOTCON 2017, Oakland, Canada (Rohatgi, 2021)) was used to extract individual stability data points from digital figures. The reliability and usability of the data extracting program has been reviewed before when applied for extraction of protein stability; proving to obtain data with high reliability, validity and usability compared to other data extracting programs (Drevon et al., 2017). To obtain data using WebPlotDigitizer, peptide stability graphs were first exported from their original manuscript into either.jpeg or.pdf files. Within the extracting software each graph axis was marked and aligned, where the x-axis referred to time and y-axis represented percentage of remaining peptide in SGF/SIF. Each data point was then accurately marked. The location of each data point (x, y), which represented time and peptide stability respectively, was used to extract stability at 30 and 120 min. These timepoints were chosen as they fall within the time that orally administered drugs would be exposed to gastric and small intestinal fluid (Awad et al., 2022; McConnell et al., 2008). This data was added to the database.

2.1.3. Peptide featurization and data preprocessing

Following data extraction, the database was structured. For each entry, the name of the peptide, the isomeric simplified molecular-input line-entry system (SMILES) notation of the peptide, the incubation environment (i.e., SGF or SIF), the percentage of drug remaining after 30 min, and the percentage of drug remaining after 120 min were inputted into the database. Where SMILES notations were not presented in the original study, they were obtained using their peptide sequence via the PepSMI tool by NovoPro (novoprolabs.com), BIPPEP-UMW (Minkiewicz et al., 2019), or obtained through the PubChem database (Kim et al., 2021). Isomeric SMILES notations were used to encode the chirality of peptides, a molecular feature known to influence GI stability (Elfgen et al., 2019; Elfgen et al., 2017).

To enable model learning the stability of peptides was binned into three categories: *Stable* (> 50 % peptide remaining at 120 min), *Unstable* (< 50 % at 30 min), and *Partly Stable* (> 50 % at 30 min and < 50 % at 120 min). Following the preliminary processing of the database, molecular featurization was carried out to represent peptides with chemically diverse features. Here, 200 physicochemical properties were calculated for each peptide with RDKit (version 2020.03.3.0) using the isomeric SMILES notations. A detailed description of the features can be found in RDKit's documentation (https://www.rdkit.org/). All feature names and the coded index used in this study was summarized in Table S1. In addition, the test environment feature was label-encoded into 1 and 0 to represent SGF and SIF, respectively. Other features were scaled within (0,1) using Equation (1):

$$X_{scaled} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (1)$$

where X represents the original value of the feature, X_{min} represents the minimum value occurring in the database for this feature and X_{max} represents the maximum. All data manipulation and pre-processing was implemented through Scikit-learn library (version 0.24.2).

2.2. Model development

2.2.1. Unsupervised data visualization

Prior to supervised ML, the dataset was explored and visualized using Seaborn library (version 0.11.1). PCA was applied to reduce the dimensions of the features into principal components, and the first two components (PC1 and PC2) were subsequently plotted to observe peptides within their feature space. The placement of peptides within their feature space allowed the analysis of structure-stability relationships by investigating the stability profiles of peptides sharing similar physicochemical properties.

Table 1
Search terms used to identify relevant data for the training dataset and the number of studies listed per term.

Search term 1	Search operator	Search term 2	Search operator	Search term 3	Number of PubMed results	Number of Google scholar results
Peptides	AND	Stability	AND	SIF	43	10,700
Peptides	AND	Stability	AND	SGF	41	9,640
Peptides	AND	Chemical Stability	AND	GI tract	10,792	55,600
Peptides	AND	Stability	AND	GI tract Fluid	33	26
Gastric stability	AND	Peptide drugs	–	–	68	32
Peptide drugs	AND	SIF stability	–	–	31	320
Peptide drugs	AND	SGF stability	–	–	23	6,740

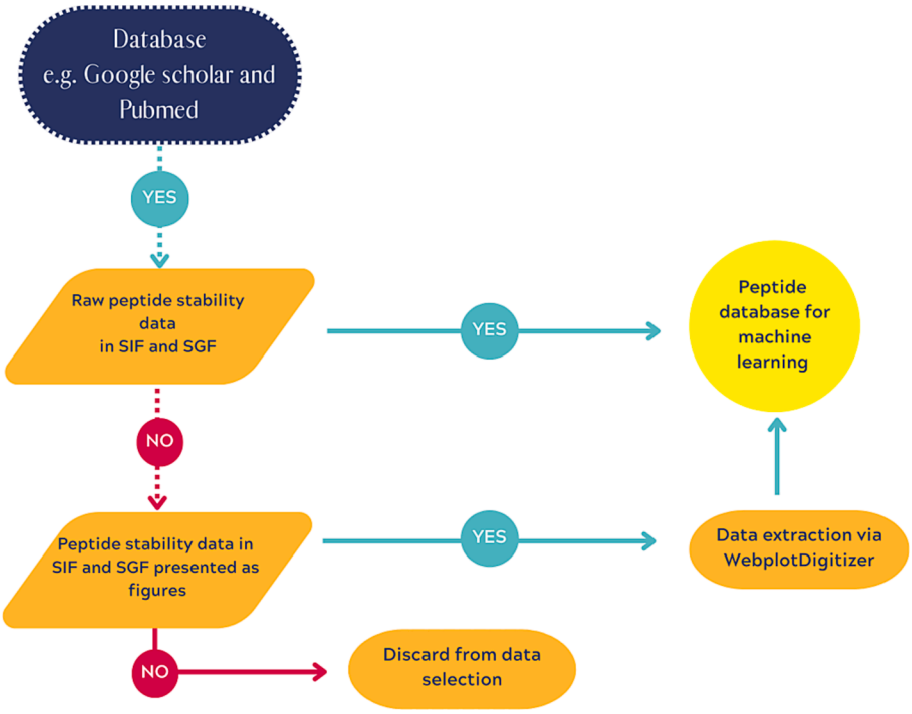

Fig. 2. Data collection flowchart for peptide stability data in SIF and SGF.

In addition, to investigate how peptides in this extracted stability database distributed in the overall therapeutic peptide space, a larger database (THPdb, accessed 07/04/22) that included all FDA-approved peptide and proteins, published by Usmani *et al.*, were plotted together using PCA (Usmani et al., 2017). In total, 239 peptides and proteins composed the THPdb. After manual removal of monoclonal antibodies and unfeaturizable molecules, 119 molecules were featurized, pre-processed and used as a representative 'peptide space'. The peptides in the training database were plotted into the THPdb peptide space using PCA, this allowed analysis of the database peptides' spread within the THPdb chemical space.

2.2.2. Supervised model benchmarking

Six different ML algorithms, known to be effective when working with smaller datasets, were developed and evaluated for prediction of peptide stability (Sugiyama, 2016; Zhang and Ling, 2018). Decision tree (DT), random forest (RF) and XGBoost were selected as representatives of tree-based models. Briefly, these tree-based models make single either-or decisions at consecutive nodes within decision trees. By combining multiple tree nodes in different manners like boosting (XGBoost) or bagging (RF), models can be better equipped to learn complicated non-linear relationships (Badillo et al., 2020; Vamathevan et al., 2019). For linear models, logistic regression with lasso (LR_Lasso) and ridge (LR_Ridge) penalty were chosen. Logistic regression is an algorithm that can convert continuous numbers (in this case the peptide physicochemical properties) through a logistic function and return categorial outputs (peptide stability) (Bishop and Nasrabadi, 2006). In addition, k-nearest neighbor (kNN) with k = 1, 2, 4 (named as kNN_1, kNN_2, and kNN_4) and support vector machine (SVM) with a linear kernel (LinearSVC) were also included. kNN is a straightforward model that classifies unlabelled samples based on their similarity to labelled samples. Linear SVM generates a linear boundary between labelled datapoints and forms predictions for new data based on their position relative to the boundary (Bishop and Nasrabadi, 2006). A more detailed explanation of the modeling methods and their applications in drug discovery and development can be found in the systematic review by Vamathevan et al. (2019). Performances of all models were benchmarked against the baseline model, which was a model that always predicted the most common class found in the training dataset. All models were trained using their default hyperparameters unless otherwise specified. Cross-validation (CV) on 5-folds with balanced accuracy and weighted f1 score (f1_weighted) as performance metrics was used to evaluate each model, as these metrics are suitable for the evaluation of unbalanced datasets (McCoubrey et al., 2021).

The equations underlying balanced accuracy and f1 score are presented in Equations (2) & (3), respectively, where true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) were calculated in a one-vs-rest manner in this multiclass classification problem. All models, except XGBoost, were implemented through the Scikit-learn library (version 0.24.2) (Pedregosa et al., 2011). The XGBoost model was based on the py-xgboost library (version 1.3.3).

$$Balanced\ accuracy = \frac{\frac{TP}{TP+FP} + \frac{TN}{TN+FN}}{2} \tag{2}$$

$$F1score = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \tag{3}$$

2.3. Feature selection and importance

Feature selection was performed in this study to examine performances of models trained on fewer, selected features. The selection process was achieved through both manual and automated selection, whereby manual selection was based on domain knowledge of peptide stability and automated feature selection screened important features automatically with a specific algorithm. Specifically, manual feature selection utilized existing domain knowledge from the previous publication by Wang *et al.* in which the colonic stability of 18 peptides was investigated (Wang et al., 2015a). Features determined as important for large intestinal stability (molecular weight, polar surface area, hydrogen bond acceptors, hydrogen bond donors, rotatable bonds, and LogP) were manually selected for the prediction of SGF and SIF stability in this study. Whereas automated feature selection with recursive feature elimination (RFE) recursively removed features that were ranked the least important by the selected model until a certain criterion was met

(e.g., an optimized accuracy where any further feature elimination had a detrimental effect on model performance) (Guyon et al., 2002). To elaborate further, in each round of feature selection, models that had the capability to evaluate feature importance (e.g., LASSO, RF, or XGBoost) produced a feature importance matrix after fitting the training data. Then, RFE was conducted by reading these matrices, eliminating three least important features, and re-fitting the model to give another round of feature important matrix. During this process, model performance was continuously monitored by 5-fold cross-validation. The feature elimination process was performed until there was no further improvement of the model performance. This was implemented through the RFECV method in the Scikit-learn library. The process of RFE was visualized as a heatmap, with the x-axis presenting the 201 features in the full feature set (1 incubation environment + 200 physicochemical properties) and the colour of the cells in the heatmap representing the importance assigned by the corresponding model on the y-axis.

The impact of feature selection techniques on models' performances was further analyzed using 5-fold CV with accuracy and f1 score as metrics. The most influential features for peptide stability were then investigated through feature importance plots and appraisal of the literature for corresponding scientific justification.

2.4. Model evaluation and interpretation

A more detailed inspection of the best-performing models was completed in addition to the evaluation with accuracy and f1 scores in model benchmarking; plotting classification confusion matrices provided an intuitive means of visualizing various types of classification errors. Here, the confusion matrices of the selected models were generated using a simple training/prediction scenario without CV. Briefly, the data of the combined dataset and two sub-datasets (SGF only and SIF only) were partitioned at an 80/20 ratio as the train and test set, respectively. The predicted results were plotted against the ground truth values as a heatmap where a lighter colour in a cell represented more instances in the corresponding category. Specifically, correct predictions, where the predicted values were the same as the ground truth values, sit on the diagonal line, whereas incorrect predictions spread over other cells in the matrix.

Model interpretation was also conducted for two models, DT and XGBoost. The prediction steps in the DT were plotted to better understand the algorithm's decision-making process. For both models, feature importances were extracted from the trained model to analyze individual features' contribution to peptide stability. However, the best model for the SGF dataset, kNN, was less interpretable by its nature and thus not analyzed further.

In addition to the classification performance analyses, selected features were investigated with unsupervised PCA to understand their contributions to the peptide stability. Like the PCA analysis for the whole dataset, PCA was also performed on the features selected as most important. Since different feature sets were used in the best performing models, PCA analysis was performed separately for both feature sets chosen for the SGF and SIF datasets. In addition, the contributions of each selected feature to the principal components were quantified and analyzed.

3. Results and discussion

3.1. Database overview

The searching queries on PubMed and Google Scholar returned approximately 874,000 publications, as indicated in Table 1. After refining the search results with specific terms and removing irrelevant articles, 16 publications were chosen for inclusion in the study (Subbaiah et al., 2019; Arif, 2018; Bertoni et al., 2019; Braga Emidio et al., 2021; Brancale et al., 2017; Cheloha et al., 2017; Claudius and Neau, 1998; Luciani et al., 2017; Ma et al., 2012; Nielsen et al., 2017; Niu et al., 2021; Pechenov et al., 2021; Wang et al., 2015a; Wang et al., 2015b; Yadav et al., 2016; Zupančič et al., 2017). From these publications a total of 109 entries of peptide stability results were extracted and formatted into the database. Exploratory analysis of the data revealed that 63/109 experiments were performed in SIF, and the remaining 46 were performed in SGF, hence, the training dataset contained greater information pertaining to SIF stability (Fig. 3). Interestingly, most experiments conducted in SIF reported peptides as unstable, whereas the majority conducted in SGF reported peptides as being stable. Prior research does affirm that peptides are more susceptible to degradation in SIF than SGF (Chen and Li, 2012; Wang et al., 2015b).

Further exploratory analysis revealed that the molecular weights of the peptides investigated ranged from 307 Da (L-glutathione) to 5778 Da (insulin), where the majority (89.9 %) of stable or partially stable peptides possessed a molecular weight below 3000 Da (Fig. 3B). Again, this finding correlates with the literature, which reports that peptides > 3000 Da are more likely to be unstable in simulated GI fluids (Chen and Li, 2012). Aside from molecular weight, the LogP feature, which describes a compound's lipophilicity, was another feature of interest, as it was previously found to positively correlate with 17 peptides' colonic stability and is recognised as an important indicator of peptide oral bioavailability (Wang et al., 2015a). However, expanding this finding to our dataset of over 60 peptides revealed no obvious correlation (Fig. 3C). The range of LogP in our database ranged from −21.5 (calcitonin) to 11.2 (lactoferrin). Notably, all physicochemical properties used here were calculated from RDKit. Thus, the absolute values were slightly different from those found in previous studies which calculated features with ChemSpider, however the values were still representative (Wang et al., 2015a; Wang et al., 2015b).

Another feature that was also initially explored was the topological polar surface area (TPSA). TPSA, defined as the total surface area of polar atoms, is correlated with molecular weights to a certain extent. The analysis revealed that GI stability favoured TPSA values below 1250 A^2 (Fig. 3D), and 97.1 % of stable or partly stable peptides had a TPSA below 1250 A^2. Indeed, research shows that peptides with lower polarity may have improved oral bioavailability (Boehm et al., 2017). Overall, these univariate analyses elucidated a degree of correlation between the stability of peptides in simulated GI fluids and defined physicochemical properties.

The findings of the exploratory data analysis inferred possible correlations between peptide stability and their molecular weight and TPSA, while the LogP feature was relatively less related. As just 2 of 200 possible physicochemical features used to describe peptides in this study, it was clear that a streamlined method for analysing GI peptide stability was needed given the high-dimensional database. Plotting and manual analysis of single features' impact on stability would be exceedingly time-consuming and would not allow appreciation of the additive effects of multiple physicochemical features. As such, ML was identified as an ideal technology due to its efficacy in modeling high-dimensional data (Castro et al., 2021; McCoubrey et al., 2022b; Ong et al., 2022).

3.2. Unsupervised modeling

Unsupervised modeling was performed using PCA, a linear modeling technique that is simple to implement and can enable visual identification of structure–activity relationships (Fig. 4) (Badillo et al., 2020; McCoubrey et al., 2022b). Visualization of all 109 datapoints, describing both SGF and SIF stability, revealed no strong relationships between peptides' physicochemical features and stability (Fig. 4A). Here, most stable peptides were clustered at PC1 values ≤ 2, however, the feature space also contained numerous unstable and several partly stable peptides, demonstrating that stability predictions for untested peptides could not be reliably formed. The clustering of peptide stability in SIF was stochastic and no discernible clusters could be observed (Fig. 4B). Further attempts to elucidate clustering for SIF samples using non-linear

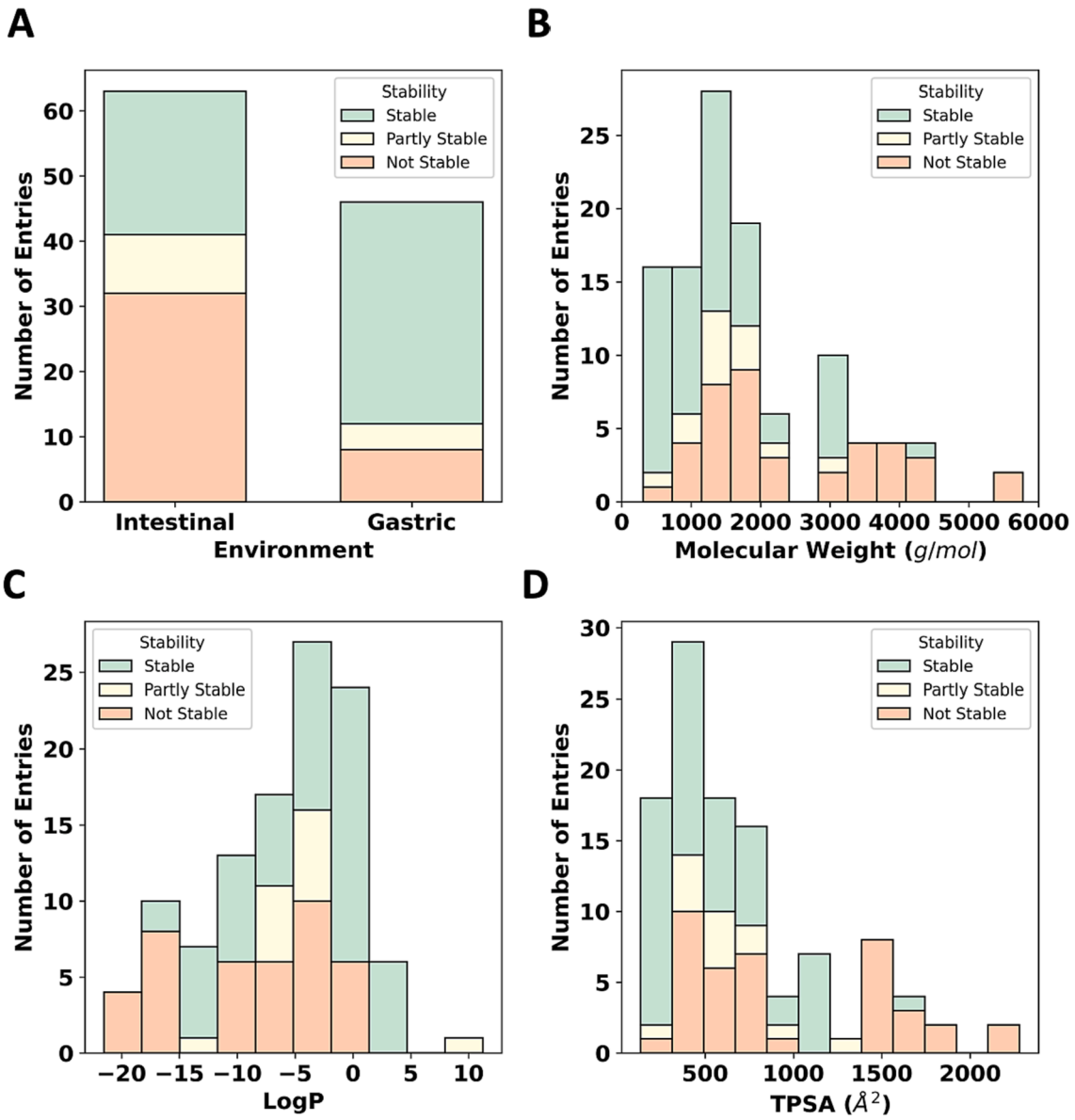

Fig. 3. Data distribution of the peptide stability database on (A) testing GI environment, (B) molecular weight, (C) LogP, and (D) TPSA (topological polar surface area).

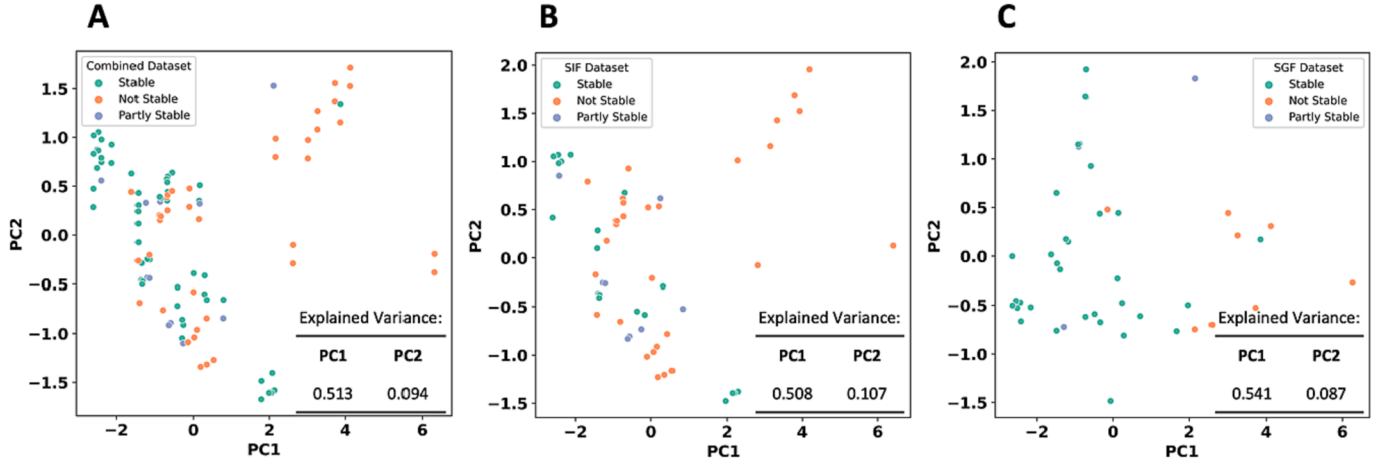

Fig. 4. PCA analysis of the database with the full feature set. (A) Full database (n = 109, SGF and SIF data). (B) Small intestinal stability database (n = 63, SIF data), (C) Gastric stability database (n = 46, SGF data).

PCA were made to no avail (Figure S1). Separating the data by their incubation medium (i.e., SGF or SIF) revealed more discernible patterns for SGF, for which it was evident that stable peptides exclusively clustered at negative PC1 values (Fig. 4C). This result gave an early forecast for promising supervised model performances on the SGF dataset since the decision boundary could already be identified on the PCA plot. It also shed light on the difficulties of modeling SIF stability. Nonetheless, unsupervised analysis revealed that PCA could be utilized to analyze peptide stability in SGF. The unsupervised learning models could offer unique benefits because the 2D plot, compared to simply outputting the predicted results, may be more visually informative to users.

PCA was also leveraged to generate a 'peptide space', composed of 119 FDA-approved therapeutic peptides and proteins, was plotted together with the peptides in the stability database (Fig. 5) (Usmani et al., 2017). The peptide space was created by conducting PCA on the physicochemical features of the peptides. Therefore, the closer the peptide markers, the more similar their physicochemical properties. There is an overlaid area between the two databases, indicating that the peptide stability data was chemically representative of peptides on the market. Less than half of the peptides in the stability database spread outside the area formed by the peptides in THPdb, though PC1 values (which accounted for most of the explained variance) were slightly different between the datasets. The properties contributing most to PC1 were HallKierAlpha, VSA_EState5, and MolLogP (positive contributors) and NumRotatableBonds, SMR_VSA10, and PEOE_VSA2 (negative contributors). This suggests that the main differences between peptides in the THPdb and stability datasets were based on these physicochemical properties.

3.3. Supervised modeling

3.3.1. Model benchmarking

Following the PCA analysis, supervised ML techniques were applied to the full and separated databases (i.e., separated SGF and SIF data). The full performances results are listed in Table S2, and the best performance for each ML technique is presented in Fig. 6. All models were significantly better at predicting peptide stability than the baseline performances (accuracy: 36.7 % for the SGF dataset, 33.3 % for the SIF dataset, and 33.3 % for the combined full dataset). For the combined dataset, the XGBoost model obtained the best accuracy at 63.4 % and an f1 score of 72.8 %. For SGF stability, kNN and LR outperformed other models with accuracies of 65.7 % and f1 scores of 83.3 %. For SIF

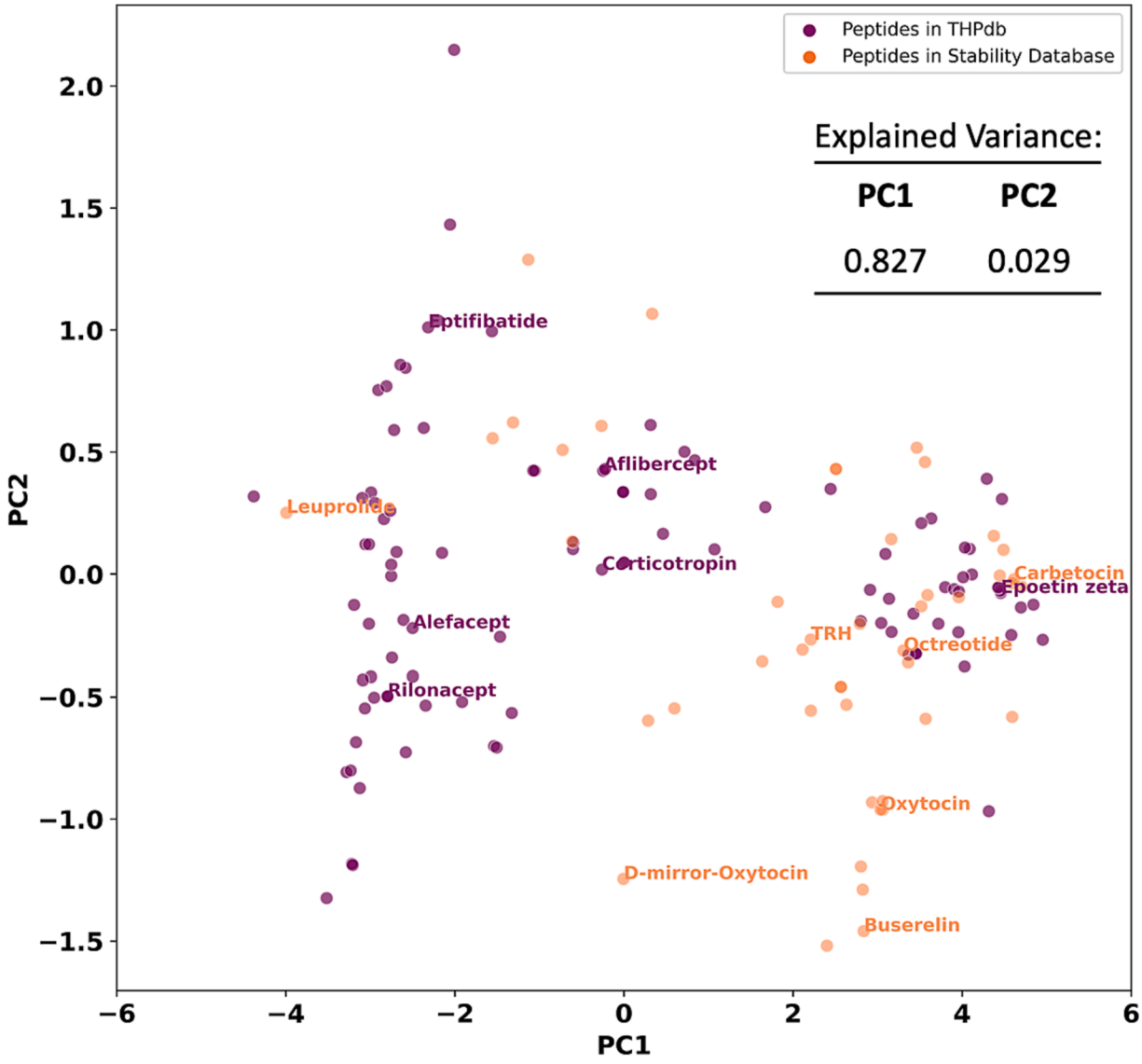

Fig. 5. Physicochemical distribution of the peptides within the stability database amongst the larger dataset of FDA-approved peptides (data from THPdb (Usmani et al., 2017)).

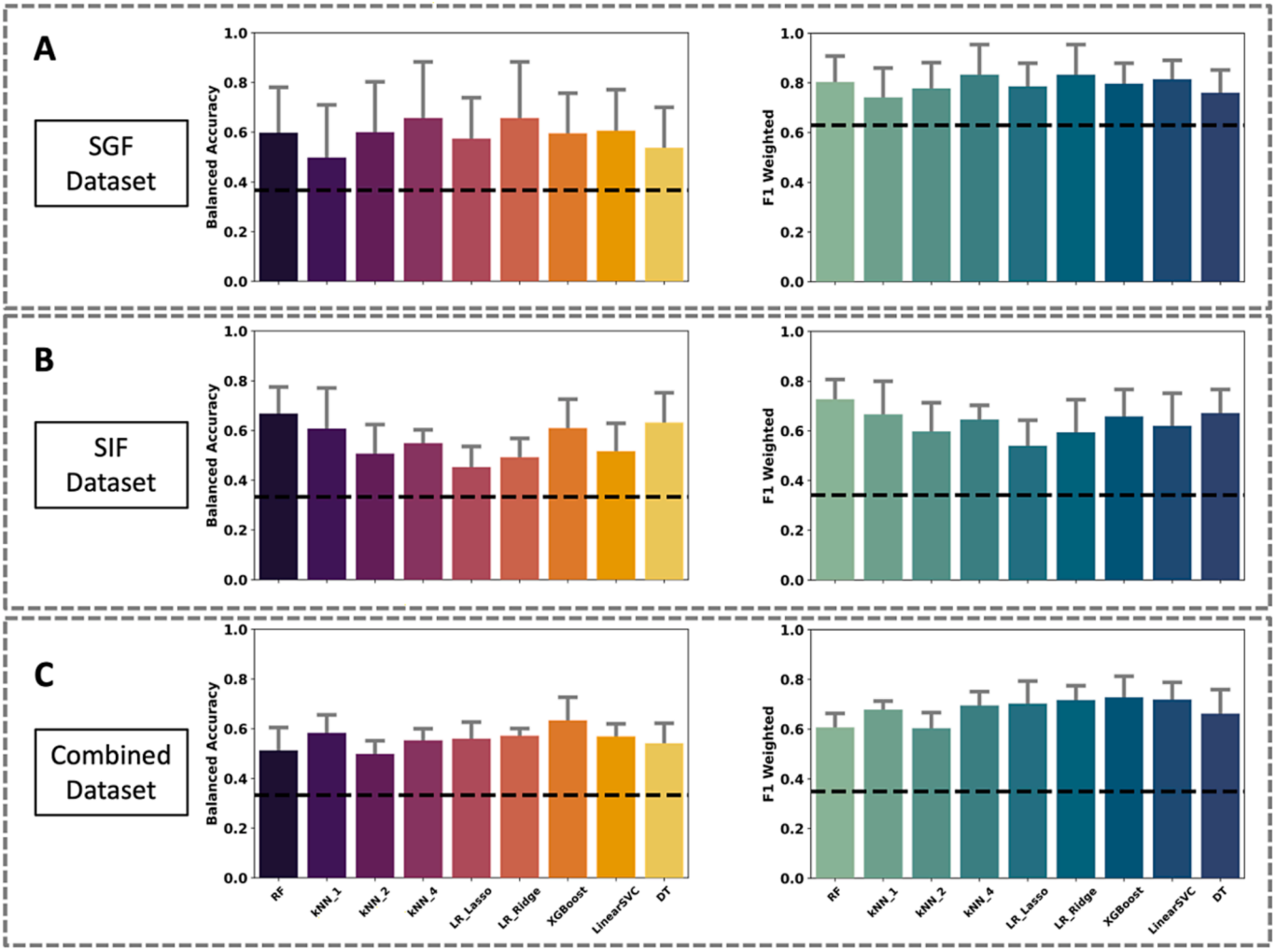

Fig. 6. Different model performances trained on SGF, SIF, and combined stability databases with the full feature set. Dashed lines represent baseline model performance, CV = 5, full results are available in Table S2.

Table 2
Features selected manually based on prior research and automated calculation (recursive feature elimination (RFE) with three different models (RF, XGBoost, LR)).

Manual		RF		XGBoost		LR	
Index	Feature name	Index	Feature name	Index	Feature name	Index	Feature name
0	Env	0	Env	0	Env	0	Env
1	Mw	55	SMR_VSA3	5	MinAbsEStateIndex	15	FpDensityMorgan1
74	TPSA	78	EState_VSA2	6	qed	18	BalabanJ
106	NumHAcceptors	82	EState_VSA6	12	MinPartialCharge	51	PEOE_VSA9
107	NumHDonors	84	EState_VSA8	25	Chi1v	59	SMR_VSA7
109	NumRotatableBonds	90	VSA_EState4	50	PEOE_VSA8	65	SlogP_VSA12
114	MolLogP			52	SMR_VSA1	70	SlogP_VSA6
				56	SMR_VSA4	77	EState_VSA11
				58	SMR_VSA6	78	EState_VSA2
				67	SlogP_VSA3	79	EState_VSA3
				76	EState_VSA10	82	EState_VSA6
				78	EState_VSA2	83	EState_VSA7
				82	EState_VSA6	90	VSA_EState4
				84	EState_VSA8	94	VSA_EState8
				85	EState_VSA9	121	fr_Ar_N
				87	VSA_EState10	133	fr_NH2
				90	VSA_EState4	134	fr_N_O
				94	VSA_EState8	138	fr_SH
						157	fr_ether
						163	fr_imidazole
						187	fr_priamide

stability, RF had the best accuracy of 66.8 % and an f1 score of 72.7 %. These results highlight that both linear and non-linear algorithms were effective in learning how peptides' physicochemical features mapped to their stability in simulated GI fluid.

3.3.2. Feature selection

To further improve the performance of models, feature selection was conducted to refine the feature set by selecting only the most influential physicochemical predicters of peptide stability. The selected features are presented in Table 2. Manual and automated selection led to the identification of various influential features that differed depending on the selection method used. Interestingly, peptide lipophilicity (represented by feature MolLogP and SLogP) was selected manually and by the XGBoost and LR automated methods, despite no obvious correlation being recognised during early dataset exploration (Section 3.1). This highlights the power of ML for identifying important factors in processes that may not be readily recognizable by simple data analysis methods (Elbadawi et al., 2021). Though prior research has identified molecular weight, TPSA, hydrogen bond acceptors, hydrogen bond donors, and rotatable bonds as important indicators of peptide colonic stability, these features were not recognized as important for SGF or SIF stability by the automated feature selection methods (Wang et al., 2015a). The feature selection process completed by RFE is presented in Figure S2, which revealed the ranking of every feature's importance for each algorithm. All three automated selection methods identified the electrotopological state, 'Estate', and corresponding van der Waals surface area, 'VSA', of atoms as highly important in predicting peptide stability (Hall and Kier, 1995; Kier and Hall, 1990). Electrotopological states can be assigned to individual atoms in a peptide, and describe atoms' electronic state as influenced by the surrounding atoms within a particular molecule (Hall et al., 1991). The higher an atom's Estate, the higher its electronegativity, thus the electronegativity of peptides can be said to influence stability in SGF and SIF.

ML analysis was repeated with the refined feature sets using the same ML techniques in Section 2; the results are presented in Fig. 7. For the combined SGF and SIF database, three of the four feature selection approaches (manual, RFE on XGBoost, and RFE on RF) successfully improved the accuracy compared to using the full 200 features, with a maximum accuracy obtained using manual feature selection paired with DT learning. Here, the accuracy and f1 scores were 65.8 % and 75.8 %, respectively, which marked improvements of 2.4 and 3.0 percentage points, respectively. The DT's decision-making process is presented in Figure S3, in which peptides' number of rotatable bonds was the first decision node, followed by MolLogP and the incubation environment (SGF/SIF). Figure S4A shows that both MolLogP and the incubation environment had relatively high importance for the model, in addition to TPSA and number of rotatable bonds. In comparison the number of hydrogen bond acceptors was deemed less important in predicting peptides' SGF/SIF stabilities.

Improvements were also observed when splitting the dataset into

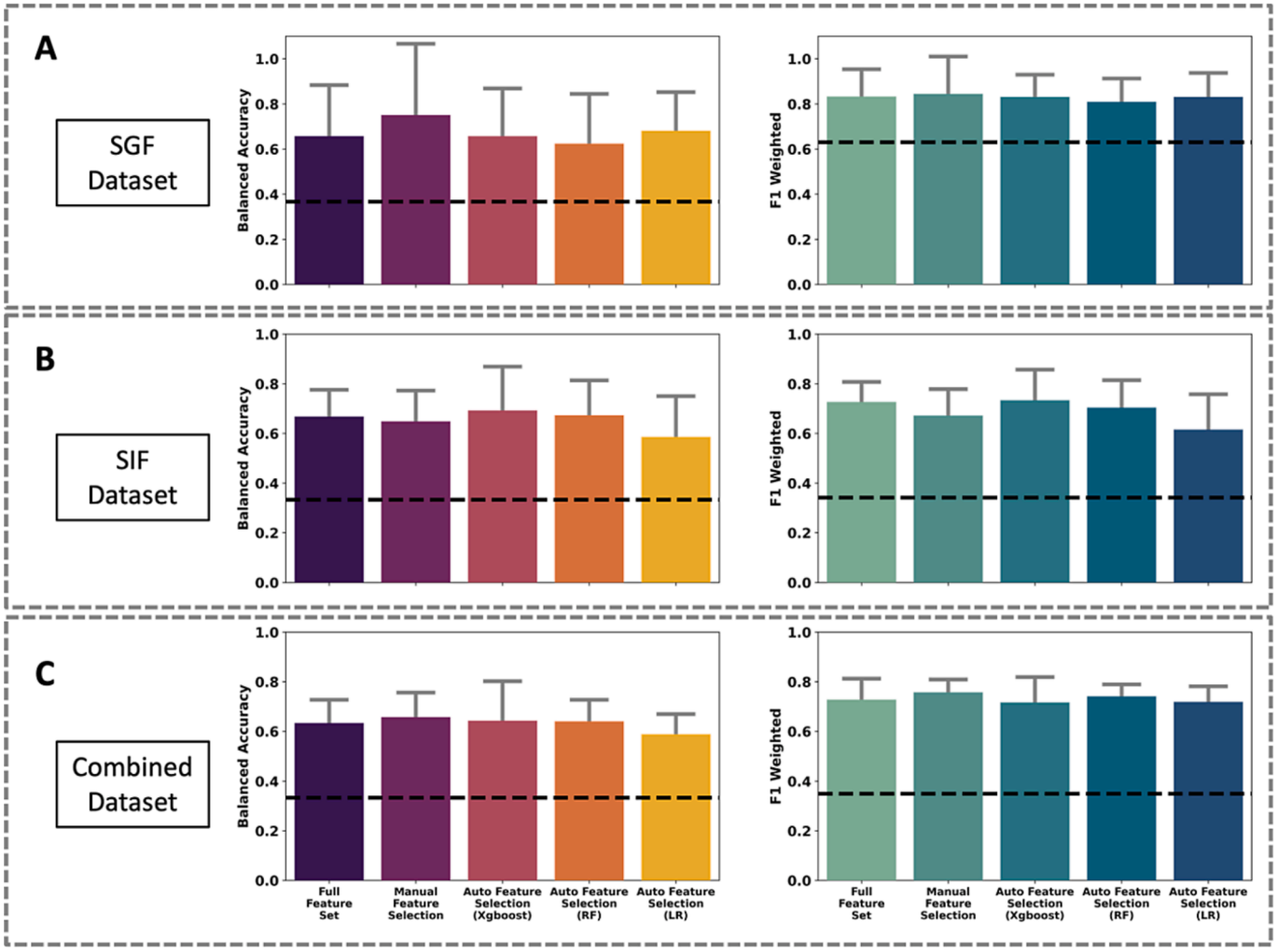

Fig. 7. Model performances on (A) SGF, (B) SIF, and (C) combined stability database with full feature set, manual feature selection, and auto feature selection. Plotted with the best results for each feature set. Dashed lines represent baseline model performance, CV = 5, full results available in Table S3.

separate SGF and SIF datasets. The best recorded accuracy and f1 scores for the SGF dataset was 75.1 % and 84.5 %, respectively, which were improvements of 9.4 and 1.2 percentage points compared with model trained on the full feature set. This was achieved by the manual feature selection approach paired with kNN learning. On the other hand, the best performing ML model for the SIF dataset was achieved by automated feature selection and XGBoost learning. The accuracy and f1 scores increased to 69.3 % and 73.4 %, an increase of 2.4 and 0.8 percentage points, respectively. The relative importance of the features within this model's learning process are presented in Figure S4B. Therefore, and despite inputting considerably fewer features, it was demonstrated that feature selection could improve model performance by limiting model training to only the most influential features.

The initial rationale of combining SGF and SIF data in the original dataset was that the mechanism of peptide breakdown or instability would be similar, and aid in learning the overall mechanisms of GI stability. However, models achieved better performances when trained on the separated SGF and SIF data, highlighting that different physicochemical features and ML techniques were required for precise prediction of peptide stability in the distinct fluids. Based on these results, it can be inferred that the features chosen via manual selection are most predictive of peptide SGF stability and those selected via automated XGBoost selection are most predictive of SIF stability. The scientific rationale for the importance of these features will be discussed in the Section 3.4.2. Dedicated models specific to SGF or SIF could be useful for new peptide therapeutics that are primarily exposed to one fluid environment; for example peptides that are exposed to gastric fluid and then rapidly absorbed across the duodenal epithelium, or enteric coated peptides that are protected from interaction with gastric fluid (Shen and Matsui, 2019).

3.4. Model evaluation and interpretation

3.4.1. Classification performance analysis

The results of the best performing models for all three datasets (combined, SGF only, SIF only) were examined. For the separate SGF and SIF datasets there was one misclassification each, which interestingly occurred between partially stable and either stable or not stable. In other words, the confusion occurred between adjacent classes. For the combined dataset, there were 3 misclassifications, where 2 misclassifications involved partially stable peptides (Fig. 8). This highlights that stability profiles lying between highly stable and highly unstable states were more challenging to predict.

3.4.2. Feature selection interpretation

Fig. 9 shows the PCA plots and feature contributions for the top principal components for the two final models. Although PCA was initially performed at an early stage (Fig. 3), here, a clearer visualization of each feature's contribution was made possible after feature selection due to greatly reduced dimensionality.

Though different ML techniques and feature sets were used for the SGF and SIF datasets, it is apparent from the feature importance that hydrophilicity/hydrophobicity of peptides was important in predicting their stability in both incubation fluids. LogP, number of hydrogen bond acceptors/donors, TPSA, Estate and peptide charge are all related to the degree to which peptides interact with the components of the aqueous fluids. It is recognised that hydrophobic peptides are shielded from interactions with aqueous solutes; for example positively charged amino acids (such as arginine) increase peptides' susceptibility to interaction with degradative enzymes, like trypsin (Kremsmayr et al., 2022). Lipophilic peptides are also afforded the benefit of increased permeability across the gastric/small intestinal epithelium (Boehm et al., 2017; Brayden et al., 2020). Thus, the inclusion of hydrophobic amino acid regions within peptides is an intelligent technique for increasing peptide stability and permeability in the GI tract.

Another strategy for imparting peptide stability in GI fluids is to increase the rigidity of peptide structures (Brayden et al., 2020; Nielsen et al., 2017). Conformational rigidity has been reported to increase peptides' GI stability and bioavailability *in vivo*, as sites vulnerable to enzymatic degradation may be shielded from access, and gut membrane permeability is increased (Nielsen et al., 2015). The kNN model developed to predict SGF stability in the present study is in agreement with the literature, as it successfully identified that the number of rotatable bonds in peptide structures is predictive of peptide stability. Decreasing the number of rotatable bonds in a peptide is an evidence-supported way to increase peptide stability in gastric fluid.

Molecular weight was also identified by the SGF model as an important indicator of peptide stability. Past research has reported that peptides $>$ 3000 Da are more likely to be hydrolyzed in SGF than smaller peptides, especially those $<$ 1000 Da (Chen and Li, 2012). Smaller peptides may show enhanced structural stability compared to larger peptides and contain fewer peptide bonds susceptible to cleavage (Wang et al., 2019). Indeed, a study examining the stability of 17 peptide therapeutics found that smaller peptides (e.g., oxytocin, desmopressin, buserelin) were significantly more stable than larger peptides (e.g., glucagon, insulin, calcitonin) during incubation with SGF for 2 h (Wang et al., 2015b). Instability was largely attributed to the higher affinity of pepsin for covalent bonds in the larger molecules, as the large peptides were significantly more stable in enzyme-free SGF. As such, researchers

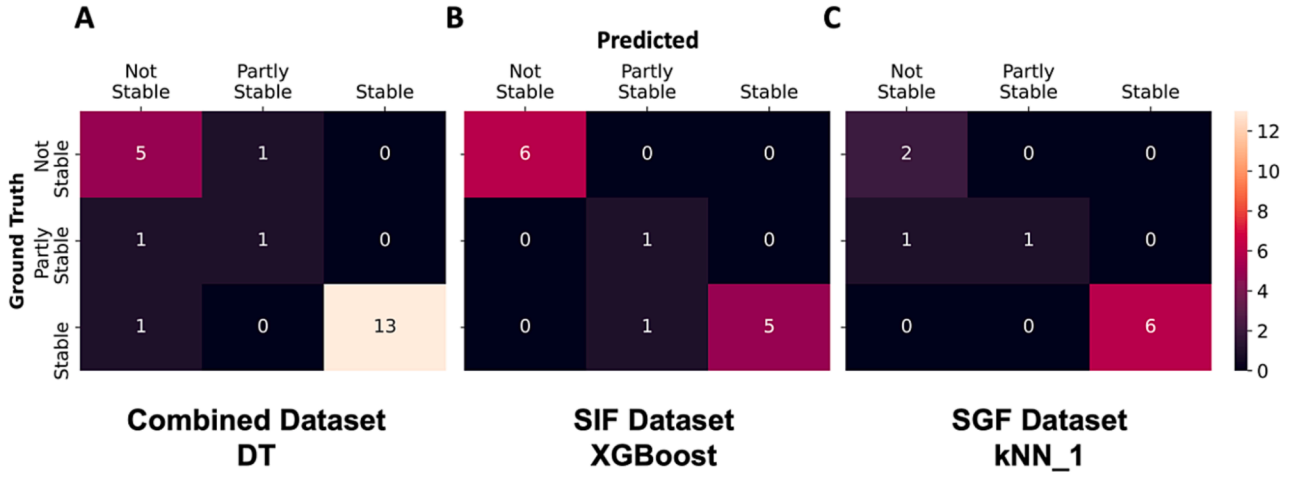

Fig. 8. Confusion matrix analysis. (A) Best performing model on combined dataset (DT, manual feature selection). (B) Best performing model on intestinal dataset (XGBoost, RFE with XGBoost feature selection). (C) Best performing model on gastric dataset (kNN_1, manual feature selection).

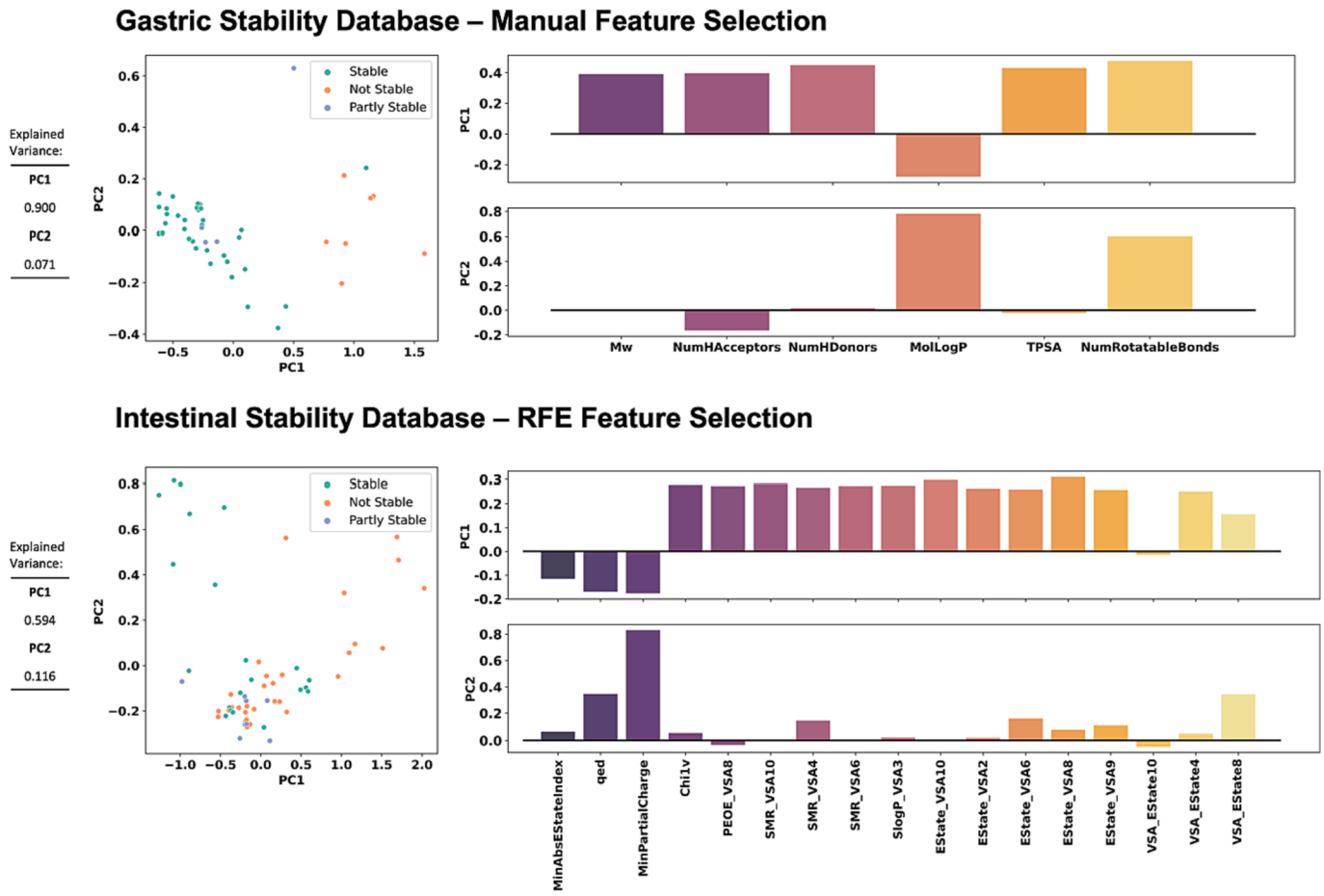

Fig. 9. PCA analysis of the gastric and intestinal stability database and feature contributions to the principal components.

developing new peptides that will be exposed to gastric fluids should consider molecular weight as an important design feature. Where peptides are required to be > 3000 Da for their therapeutic effect, enteric coats that shield the active molecules from pepsin could be utilized (Awad et al., 2022).

The use of USP simulated GI fluids to study the behaviour of oral pharmaceutical formulations is common practice, therefore our model provides a means of predicting peptide stability in these media. With access to relevant data at the required quantities for ML, the generalizability of our model for *in vivo* conditions could be improved by considering peptide stability in more physiologically relevant fluids, such as those containing bile salts and small intestinal microbiota (Dening et al., 2021; McCoubrey et al., 2021). Another interesting ML task could be to predict the epithelial permeability of peptide drugs. Paired with prediction of GI stability, estimation of peptides' intestinal permeability could facilitate calculation of their expected oral bioavailability.

4. Conclusion

In this study ML was applied to predict the stability of therapeutic peptides in SGF and SIF, using a training dataset of 109 peptide stability results extracted from the literature. Initial dataset exploration revealed that peptides with lower molecular weights (< 3000 Da) were more likely to be stable in both SGF and SIF. Further, TPSA values < 1250 A^2 were predictive of stability in both media. PCA clustering of all 109 incubation results or SIF data alone did not lead to distinct relationships, however discernible peptide structure-stability patterns did emerge when clustering the SGF data alone. The best performing supervised ML models consisted of a kNN model trained on manually selected features for prediction of peptide stability in SGF, and an XGBoost model trained on automatically selected features for prediction of peptide stability in SIF. The accuracies of these models in 5-fold cross-validation were 75.1 % (kNN model, for SGF) and 69.3 % (XGBoost model, for SIF); the f1 scores were 84.5 % (kNN model, for SGF) and 73.4 % (XGBoost model, for SIF). Feature importance revealed that physicochemical properties pertaining to peptide molecular weight, hydrophobicity/hydrophilicity, and conformational flexibility were most influential in predicting peptide stability in the GI fluids. Importantly, these features agree with findings from preclinical and human studies, and provide evidence-assured strategies for researchers working to develop novel orally administrable peptide therapeutics. The models developed in this study have been made available for predicting the stability of untested peptides in SGF and SIF, and may be used to digitally screen the suitability of peptide drugs for oral administration.

CRediT authorship contribution statement

Fanjin Wang: Methodology, Software, Formal analysis, Investigation, Writing – original draft, Writing – review & editing, Visualization. **Nannapat Sangfuang:** Methodology, Software, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Laura E. McCoubrey:** Methodology, Software, Data curation, Writing – original draft, Writing – review & editing, Project administration. **Vipul Yadav:** Conceptualization, Methodology, Writing – review & editing, Supervision, Funding acquisition. **Moe Elbadawi:** Conceptualization, Methodology, Software, Formal analysis, Writing – review & editing, Project administration. **Mine Orlu:** Supervision, Funding acquisition. **Simon Gaisford:** Supervision, Funding acquisition. **Abdul W. Basit:** Conceptualization, Methodology, Writing –

review & editing, Supervision, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

A hyperlink to the GitHub repository has been made available in-text.

Acknowledgements

The authors acknowledge Intract Pharma Ltd. and The Engineering and Physical Sciences Research Council grants [EP/S023054/1; EP/S009000/1] to UCL School of Pharmacy for funding this work. BioRender is also acknowledged for its use in designing the graphical abstract.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ijpharm.2023.122643.

References

Abramson, A., Frederiksen, M.R., Vegge, A., Jensen, B., Poulsen, M., Mouridsen, B., Jespersen, M.O., Kirk, R.K., Windum, J., Hubálek, F., Water, J.J., Fels, J., Gunnarsson, S.B., Bohr, A., Straarup, E.M., Ley, M.W.H., Lu, X., Wainer, J., Collins, J., Tamang, S., Ishida, K., Hayward, A., Herskind, P., Buckley, S.T., Roxhed, N., Langer, R., Rahbek, U., Traverso, G., 2022. Oral delivery of systemic monoclonal antibodies, peptides and small molecules using gastric auto-injectors. Nat. Biotechnol. 40, 103–109.

Ahmed, T., Sun, X., Udenigwe, C.C., 2022. Role of structural properties of bioactive peptides in their stability during simulated gastrointestinal digestion: A systematic review. Trends Food Sci. Technol. 120, 265–273.

Arif, M.I., 2018. Engineering amidases for peptide C-terminal modification. University of Groningen (Netherlands). Thesis.

Awad, A., Madla, C.M., McCoubrey, L.E., Ferraro, F., Gavins, F.K.H., Buanz, A., Gaisford, S., Orlu, M., Siepmann, F., Siepmann, J., Basit, A.W., 2022. Clinical translation of advanced colonic drug delivery technologies. Adv. Drug Deliv. Rev. 181, 114076.

Badillo, S., Banfai, B., Birzele, F., Davydov, I.I., Hutchinson, L., Kam-Thong, T., Siebourg-Polster, J., Steiert, B., Zhang, J.D., 2020. An Introduction to Machine Learning. Clin. Pharmacol. Ther. 107, 871–885.

Bell, J., 2022. What Is Machine Learning?, Machine Learning and the City, pp. 207-216.

Bertoni, S., Albertini, B., Facchini, C., Prata, C., Passerini, N., 2019. Glutathione-loaded solid lipid microparticles as innovative delivery system for oral antioxidant therapy. Pharmaceutics 11, 364.

Bishop, C.M., Nasrabadi, N.M., 2006. Pattern recognition and machine learning. Springer.

Boehm, M., Beaumont, K., Jones, R., Kalgutkar, A.S., Zhang, L., Atkinson, K., Bai, G., Brown, J.A., Eng, H., Goetz, G.H., Holder, B.R., Khunte, B., Lazzaro, S., Limberakis, C., Ryu, S., Shapiro, M.J., Tylaska, L., Yan, J., Turner, R., Leung, S.S.F., Ramaseshan, M., Price, D.A., Liras, S., Jacobson, M.P., Earp, D.J., Lokey, R.S., Mathiowetz, A.M., Menhaji-Klotz, E., 2017. Discovery of Potent and Orally Bioavailable Macrocyclic Peptide-Peptoid Hybrid CXCR7 Modulators. J. Med. Chem. 60, 9653–9663.

Bou-Chacra, N., Melo, K.J.C., Morales, I.A.C., Stippler, E.S., Kesisoglou, F., Yazdanian, M., Löbenberg, R., 2017. Evolution of Choice of Solubility and Dissolution Media After Two Decades of Biopharmaceutical Classification System. AAPS J. 19, 989–1001.

Boye, K.S., Matza, L.S., Walter, K.N., Van Brunt, K., Palsgrove, A.C., Tynan, A., 2011. Utilities and disutilities for attributes of injectable treatments for type 2 diabetes. Eur. J. Health Econ. 12, 219–230.

Braga Emidio, N., Tran, H.N., Andersson, A., Dawson, P.E., Albericio, F., Vetter, I., Muttenthaler, M., 2021. Improving the gastrointestinal stability of linaclotide. J. Med. Chem. 64, 8384–8390.

Brancale, A., Shailubhai, K., Ferla, S., Ricci, A., Bassetto, M., Jacob, G.S., 2017. Therapeutically targeting guanylate cyclase-C: computational modeling of plecanatide, a uroguanylin analog. Pharmacol. Res. Perspect. 5, e00295.

Brayden, D.J., Hill, T.A., Fairlie, D.P., Maher, S., Mrsny, R.J., 2020. Systemic delivery of peptides by the oral route: Formulation and medicinal chemistry approaches. Adv. Drug Deliv. Rev. 157, 2–36.

Camela, E., Ocampo-Garza, S.S., Cinelli, E., Villani, A., Fabbrocini, G., Megna, M., 2021. Therapeutic update of biologics and small molecules for scalp psoriasis: a systematic review. Dermatol. Ther. 34, e14857.

Castro, B.M., Elbadawi, M., Ong, J.J., Pollard, T., Song, Z., Gaisford, S., Perez, G., Basit, A.W., Cabalar, P., Goyanes, A., 2021. Machine learning applied to over 900 3D printed drug delivery systems. J Control Release.

Chandrasekaran, B., Abed, S.N., Al-Attraqchi, O., Kuche, K., Tekade, R.K., 2018. Chapter 21 - Computer-Aided Prediction of Pharmacokinetic (ADMET) Properties. In: Tekade, R.K. (Ed.), Dosage Form Design Parameters. Academic Press, pp. 731–755.

Cheloha, R.W., Chen, B., Kumar, N.N., Watanabe, T., Thorne, R.G., Li, L., Gardella, T.J., Gellman, S.H., 2017. Development of potent, protease-resistant agonists of the parathyroid hormone receptor with broad β residue distribution. J. Med. Chem. 60, 8816–8833.

Chen, M., Li, B., 2012. The effect of molecular weights on the survivability of casein-derived antioxidant peptides after the simulated gastrointestinal digestion. Innov. Food Sci. Emerg. Technol. 16, 341–348.

Claudius, J.S., Neau, S.H., 1998. The solution stability of vancomycin in the presence and absence of sodium carboxymethyl starch. Int. J. Pharm. 168, 41–48.

Dening, T.J., Douglas, J.T., Hageman, M.J., 2021. Do Macrocyclic Peptide Drugs Interact with Bile Salts under Simulated Gastrointestinal Conditions? Mol. Pharm. 18, 3086–3098.

Drevon, D., Fursa, S.R., Malcolm, A.L., 2017. Intercoder Reliability and Validity of WebPlotDigitizer in Extracting Graphed Data. Behav. Modif. 41, 323–339.

Drucker, D.J., 2020. Advances in oral peptide therapeutics. Nat. Rev. Drug Discov. 19, 277–289.

Elbadawi, M., McCoubrey, L.E., Gavins, F.K.H., Jie Ong, J., Goyanes, A., Gaisford, S., Basit, A.W., 2021. Harnessing Artificial Intelligence for the Next Generation of 3D Printed Medicines. Adv. Drug Deliv. Rev. 175, 113805.

Elfgen, A., Santiago-Schubel, B., Gremer, L., Kutzsche, J., Willbold, D., 2017. Surprisingly high stability of the Abeta oligomer eliminating all-d-enantiomeric peptide D3 in media simulating the route of orally administered drugs. Eur. J. Pharm. Sci. 107, 203–207.

Elfgen, A., Hupert, M., Bochinsky, K., Tusche, M., de San, G., Roman Martin, E., Gering, I., Sacchi, S., Pollegioni, L., Huesgen, P.F., Hartmann, R., Santiago-Schubel, B., Kutzsche, J., Willbold, D., 2019. Metabolic resistance of the D-peptide RD2 developed for direct elimination of amyloid-beta oligomers. Sci. Rep. 9, 5715.

Forbes, J., Krishnamurthy, K., 2022. Biochemistry, Peptide, StatPearls. StatPearls Publishing Copyright © 2022, StatPearls Publishing LLC., Treasure Island (FL).

Gao, Y., Gesenberg, C., Zheng, W., 2017. Chapter 17 - Oral Formulations for Preclinical Studies: Principle, Design, and Development Considerations. In: Qiu, Y., Chen, Y., Zhang, G.G.Z., Yu, L., Mantri, R.V. (Eds.), Developing Solid Oral Dosage Forms, Second Edition. Academic Press, Boston, pp. 455–495.

Gao, W., Mahajan, S.P., Sulam, J., Gray, J.J., 2020. Deep Learning in Protein Structural Modeling and Design. Patterns 1, 100142.

Guyon, I., Weston, J., Barnhill, S., Vapnik, V., 2002. Gene selection for cancer classification using support vector machines. Mach. Learn. 46, 389–422.

Hall, L.H., Kier, L.B., 1995. Electrotopological State Indices for Atom Types: A Novel Combination of Electronic, Topological, and Valence State Information. J. Chem. Inf. Comput. Sci. 35, 1039–1045.

Hall, L.H., Mohney, B., Kier, L.B., 1991. The Electrotopological State: An Atom Index for QSAR. Quant. Struct.-Act. Relat. 10, 43–51.

Hatton, G.B., Yadav, V., Basit, A.W., Merchant, H.A., 2015. Animal Farm: Considerations in Animal Gastrointestinal Physiology and Relevance to Drug Delivery in Humans. J. Pharm. Sci. 104, 2747–2776.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S.A.A., Ballard, A.J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A.W., Kavukcuoglu, K., Kohli, P., Hassabis, D., 2021. Highly accurate protein structure prediction with AlphaFold. Nature.

Kier, L.B., Hall, L.H., 1990. An Electrotopological-State Index for Atoms in Molecules. Pharm. Res. 7, 801–807.

Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B.A., Thiessen, P.A., Yu, B., Zaslavsky, L., Zhang, J., Bolton, E.E., 2021. PubChem in 2021: new data content and improved web interfaces. Nucl. Acids Res. 49, D1388–D1395.

Klepach, A., Tran, H., Ahmad Mohammed, F., ElSayed, M.E.H., 2022. Characterization and impact of peptide physicochemical properties on oral and subcutaneous delivery. Adv. Drug Deliv. Rev. 186, 114322.

Kremsmayr, T., Aljnabi, A., Blanco-Canosa, J.B., Tran, H.N.T., Emidio, N.B., Muttenthaler, M., 2022. On the Utility of Chemical Strategies to Improve Peptide Gut Stability. J. Med. Chem.

Lasa, J.S., Olivera, P.A., Danese, S., Peyrin-Biroulet, L., 2022. Efficacy and safety of biologics and small molecule drugs for patients with moderate-to-severe ulcerative colitis: a systematic review and network meta-analysis. Lancet Gastroenterol. Hepatol. 7, 161–170.

Lau, J.L., Dunn, M.K., 2018. Therapeutic peptides: Historical perspectives, current development trends, and future directions. Bioorg. Med. Chem. 26, 2700–2707.

Lohman, R.-J., Nielsen, D.S., Kok, W.M., Hoang, H.N., Hill, T.A., Fairlie, D.P., 2019. Mirror image pairs of cyclic hexapeptides have different oral bioavailabilities and metabolic stabilities. Chem. Commun. 55, 13362–13365.

Luciani, P., Estella-Hermoso de Mendoza, A., Casalini, T., Lang, S., Atrott, K., Spalinger, M.R., Pratsinis, A., Sobek, J., Frey-Wagner, I., Schumacher, J., Leroux, J.-C., Rogler, G., 2017. Gastroresistant oral peptide for fluorescence imaging of colonic inflammation. J. Control. Release 262, 118–126.

Ma, B., Yin, C., Yang, D., Lin, G., 2012. Effect of structural modification on the gastrointestinal stability and hepatic metabolism of α-aminoxy peptides. Amino Acids 43, 2073–2085.
Makurvet, F.D., 2021. Biologics vs. small molecules: Drug costs and patient access. Medicine. Drug Discov. 9, 100075.
McConnell, E.L., Fadda, H.M., Basit, A.W., 2008. Gut instincts: explorations in intestinal physiology and drug delivery. Int. J. Pharm. 364, 213–226.
McCoubrey, L.E., Thomaidou, S., Elbadawi, M., Gaisford, S., Orlu, M., Basit, A.W., 2021. Machine Learning Predicts Drug Metabolism and Bioaccumulation by Intestinal Microbiota. Pharmaceutics 13.
McCoubrey, L.E., Favaron, A., Awad, A., Orlu, M., Gaisford, S., Basit, A.W., 2022a. Colonic drug delivery: Formulating the next generation of colon-targeted therapeutics. J. Control. Release 353, 1107–1126.
McCoubrey, L.E., Seegobin, N., Elbadawi, M., Hu, Y., Orlu, M., Gaisford, S., Basit, A.W., 2022b. Active Machine Learning for Formulation of Precision Probiotics. Int. J. Pharm. 121568.
Minkiewicz, P., Iwaniak, A., Darewicz, M., 2019. BIOPEP-UWM Database of Bioactive Peptides: Current Opportunities. Int. J. Mol. Sci. 20.
Narayanan, H., Dingfelder, F., Butté, A., Lorenzen, N., Sokolov, M., Arosio, P., 2021. Machine Learning for Biologics: Opportunities for Protein Engineering, Developability, and Formulation. Trends Pharmacol. Sci. 42, 151–165.
Nielsen, D.S., Lohman, R.-J., Hoang, H.N., Hill, T.A., Jones, A., Lucke, A.J., Fairlie, D.P., 2015. Flexibility versus Rigidity for Orally Bioavailable Cyclic Hexapeptides. Chembiochem 16, 2289–2293.
Nielsen, D.S., Shepherd, N.E., Xu, W., Lucke, A.J., Stoermer, M.J., Fairlie, D.P., 2017. Orally Absorbed Cyclic Peptides. Chem. Rev. 117, 8094–8128.
Niu, Z., Thielen, I., Loveday, S.M., Singh, H., 2021. Emulsions Stabilised by Polyethylene Glycol (PEG) 40 Stearate and Lactoferrin for Protection of Lactoferrin during In Vitro Digestion. Food Biophys. 16, 40–47.
Ong, J.J., Castro, B.M., Gaisford, S., Cabalar, P., Basit, A.W., Pérez, G., Goyanes, A., 2022. Accelerating 3D printing of pharmaceutical products using machine learning. Int. J. Pharmaceutics: X 4, 100120.
Pechenov, S., Revell, J., Will, S., Naylor, J., Tyagi, P., Patel, C., Liang, L., Tseng, L., Huang, Y., Rosenbaum, A.I., 2021. Development of an orally delivered GLP-1 receptor agonist through peptide engineering and drug delivery to treat chronic disease. Sci. Rep. 11, 1–15.
Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Andreas Müller, Joel Nothman, Gilles Louppe, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, Duchesnay, É., 2011. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 2825-2830.
Rohatgi, A., 2021. WebPlotDigitizer. Austin, Texas, USA.
Shen, W., Matsui, T., 2019. Intestinal absorption of small peptides: a review. Int. J. Food Sci. Technol. 54, 1942–1948.
Subbaiah, A.M.M., Mandlekar, S., Desikan, S., Ramar, T., Subramani, L., Annadurai, M., Desai, S.D., Sinha, S., Jenkins, S.M., Krystal, M.R., 2019. Design, synthesis, and pharmacokinetic evaluation of phosphate and amino acid ester prodrugs for improving the oral bioavailability of the HIV-1 protease inhibitor atazanavir. J. Med. Chem. 62, 3553–3574.
Sugiyama, M., 2016. Chapter 21 - Learning Models. In: Sugiyama, M. (Ed.), Introduction to Statistical Machine Learning. Morgan Kaufmann, Boston, pp. 237–244.
Usmani, S.S., Bedi, G., Samuel, J.S., Singh, S., Kalra, S., Kumar, P., Ahuja, A.A., Sharma, M., Gautam, A., Raghava, G.P.S., 2017. THPdb: Database of FDA-approved peptide and protein therapeutics. PLoS One 12, e0181748.
Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., Li, B., Madabhushi, A., Shah, P., Spitzer, M., Zhao, S., 2019. Applications of machine learning in drug discovery and development. Nat. Rev. Drug Discov. 18, 463–477.
Wang, B., Xie, N., Li, B., 2019. Influence of peptide characteristics on their stability, intestinal transport, and in vitro bioavailability: A review. J. Food Biochem. 43, e12571.
Wang, J., Yadav, V., Smart, A.L., Tajiri, S., Basit, A.W., 2015a. Stability of peptide drugs in the colon. Eur. J. Pharm. Sci. 78, 31–36.
Wang, J., Yadav, V., Smart, A.L., Tajiri, S., Basit, A.W., 2015b. Toward Oral Delivery of Biopharmaceuticals: An Assessment of the Gastrointestinal Stability of 17 Peptide Drugs. Mol. Pharm. 12, 966–973.
Whitcomb, D.C., Lowe, M.E., 2007. Human Pancreatic Digestive Enzymes. Dig. Dis. Sci. 52, 1–17.
Wicke, N., Bedford, M.R., Howarth, M., 2021. Gastrobodies are engineered antibody mimetics resilient to pepsin and hydrochloric acid. Commun. Biol. 4, 960.
Yadav, V., Varum, F., Bravo, R., Furrer, E., Basit, A.W., 2016. Gastrointestinal stability of therapeutic anti-TNF alpha IgG1 monoclonal antibodies. Int. J. Pharm. 502, 181–187.
Zhang, Y., Ling, C., 2018. A strategy to apply machine learning to small datasets in materials science. npj Comput. Mater. 4, 25.
Zhang, W., Michalowski, C.B., Beloqui, A., 2021. Oral Delivery of Biologics in Inflammatory Bowel Disease Treatment. Front. Bioeng. Biotechnol. 9, 675194.
Zizzari, A.T., Pliatsika, D., Gall, F.M., Fischer, T., Riedl, R., 2021. New perspectives in oral peptide delivery. Drug Discov. Today 26, 1097–1105.
Zupančič, O., Rohrer, J., Thanh Lam, H., Grießinger, J.A., Bernkop-Schnürch, A., 2017. Development and in vitro characterization of self-emulsifying drug delivery system (SEDDS) for oral opioid peptide delivery. Drug Dev. Ind. Pharm. 43, 1694–1702.