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

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

A R T I C L E   I N F O

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

A B S T R A C T

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 compar-}

<b>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 Pharmacopoeia.</b>

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

0378-5173/© 2023 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).
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 (Brauden 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 featureization 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 (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.

#### 2.1.2. 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 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).

### 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

<table>
<thead>
<tr>
<th>Search term 1</th>
<th>Search operator</th>
<th>Search term 2</th>
<th>Search operator</th>
<th>Search term 3</th>
<th>Number of PubMed results</th>
<th>Number of Google scholar results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peptides</td>
<td>AND</td>
<td>Stability</td>
<td>AND</td>
<td>SIF</td>
<td>43</td>
<td>10,700</td>
</tr>
<tr>
<td>Peptides</td>
<td>AND</td>
<td>Stability</td>
<td>AND</td>
<td>SGF</td>
<td>41</td>
<td>9,640</td>
</tr>
<tr>
<td>Peptides</td>
<td>AND</td>
<td>Chemical Stability</td>
<td>AND</td>
<td>GI tract</td>
<td>10,792</td>
<td>55,600</td>
</tr>
<tr>
<td>Peptides</td>
<td>AND</td>
<td>Stability</td>
<td>AND</td>
<td>GI tract fluid</td>
<td>33</td>
<td>26</td>
</tr>
<tr>
<td>Gastric stability</td>
<td>AND</td>
<td>Peptide drugs</td>
<td>–</td>
<td>–</td>
<td>68</td>
<td>22</td>
</tr>
<tr>
<td>Peptide drugs</td>
<td>AND</td>
<td>SIF stability</td>
<td>–</td>
<td>–</td>
<td>31</td>
<td>320</td>
</tr>
<tr>
<td>Peptide drugs</td>
<td>AND</td>
<td>SFG stability</td>
<td>–</td>
<td>–</td>
<td>23</td>
<td>6,740</td>
</tr>
</tbody>
</table>
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 categorical 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 data points 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 ($F_1$ 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 $F_1$ 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).

$$\text{Balanced accuracy} = \frac{TP + TN}{2}$$

$$\text{F1 score} = \frac{TP}{TP + \frac{1}{2}(FP + FN)}$$

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.
3. Results and discussion

Each selected feature to the principal components were quantified and chosen for the SGF and SIF datasets. In addition, the contributions of models, PCA analysis was performed separately for both feature sets (i.e., additive effects of multiple physicochemical features). As such, ML was used to predict peptide stability needs (Chen and Li, 2012; 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 Å² (Fig. 3D), and 97.1 % of stable or partly stable peptides had a TPSA below 1250 Å². 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. 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)).

<table>
<thead>
<tr>
<th></th>
<th>Manual RF</th>
<th>Feature name</th>
<th>Index</th>
<th></th>
<th>Feature name</th>
<th>Index</th>
<th></th>
<th>Feature name</th>
<th>Index</th>
<th></th>
<th>Feature name</th>
<th>Index</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Env</td>
<td>Env</td>
<td>0</td>
<td></td>
<td>Env</td>
<td>0</td>
<td></td>
<td>Env</td>
<td>0</td>
<td></td>
<td>Env</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>Mw</td>
<td>SMR_VSA3</td>
<td>55</td>
<td></td>
<td>MinAbsEStateIndex</td>
<td>5</td>
<td></td>
<td>FpDensityMorgan1</td>
<td>15</td>
<td></td>
<td>SlogP_VSA12</td>
<td>65</td>
</tr>
<tr>
<td>74</td>
<td>TPSA</td>
<td>EState_VSA2</td>
<td>78</td>
<td></td>
<td>qed</td>
<td>6</td>
<td></td>
<td>PEOE_VSA9</td>
<td>51</td>
<td></td>
<td>fr_Ar_N</td>
<td>121</td>
</tr>
<tr>
<td>106</td>
<td>NumHAcceptors</td>
<td>EState_VSA6</td>
<td>82</td>
<td></td>
<td>MinPartialCharge</td>
<td>12</td>
<td></td>
<td>SlogP_VSA12</td>
<td>65</td>
<td></td>
<td>fr_Ar_N</td>
<td>121</td>
</tr>
<tr>
<td>107</td>
<td>NumHDonors</td>
<td>EState_VSA8</td>
<td>84</td>
<td></td>
<td>Chi1v</td>
<td>25</td>
<td></td>
<td>SlogP_VSA12</td>
<td>65</td>
<td></td>
<td>fr_Ar_N</td>
<td>121</td>
</tr>
<tr>
<td>109</td>
<td>NumRotatableBonds</td>
<td>VSA_EState4</td>
<td>90</td>
<td></td>
<td>PEOE_VSA8</td>
<td>50</td>
<td></td>
<td>SlogP_VSA12</td>
<td>65</td>
<td></td>
<td>fr_Ar_N</td>
<td>121</td>
</tr>
<tr>
<td>114</td>
<td>MolLogP</td>
<td>VSA_EState4</td>
<td>90</td>
<td></td>
<td>PEOE_VSA8</td>
<td>50</td>
<td></td>
<td>SlogP_VSA12</td>
<td>65</td>
<td></td>
<td>fr_Ar_N</td>
<td>121</td>
</tr>
</tbody>
</table>
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 predictors 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.](image-url)
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...
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 Å² 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/5023054/1; EP/ S009900/1] 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


