Predviđanje samosastavljanja peptida zasnovano na sklonosti agregaciji i sekvencijalnim značajkama

Žužić, Lucija

Master's thesis / Diplomski rad

2023

Degree Grantor / Ustanova koja je dodijelila akademski / stručni stupanj: **University of Rijeka, Faculty of Engineering / Sveučilište u Rijeci, Tehnički fakultet**

Permanent link / Trajna poveznica: https://urn.nsk.hr/urn:nbn:hr:190:971248

SVEUČILIŠTE U RIJECI
TEHNIČKI FAKULTET
Diplomski sveučilišni studij računarstva

Diplomski rad

Predviđanje samosastavljanja peptida zasnovano na sklonosti agregaciji i sekvencijalnim značajkama

Rijeka, srpanj 2023.

Lucija Žužić
0069085398

SVEUČILIŠTE U RIJECI
TEHNIČKI FAKULTET
Diplomski sveučilišni studij računarstva

Diplomski rad

Predviđanje samosastavljanja peptida zasnovano na sklonosti agregaciji i sekvencijalnim značajkama

Mentor: doc. dr. sc. Goran Mauša

Rijeka, srpanj 2023.

Lucija Žužić
0069085398

Umjesto ove stranice umetnuti zadatak za završni ili diplomski rad

Izjava o samostalnoj izradi rada

Izjavljujem da sam samostalno izradila ovaj rad.

Rijeka, srpanj 2023.

_____
Lucija Žužić

Zahvala

Zahvaljujem dr. sc. Goranu Mauši što mi je bio mentor na izbornom projektu, na završnom i na diplomskom radu. Zahvaljujem mu se na uloženom trudu u zajednički rad i savjetima kako da ga prezentiram na najbolji mogući način.

Zahvaljujem asist. Marku Njirjaku na dijeljenju znanja o konfiguraciji neuronskih mreža i na spajanju modela neuronske mreže s genetskim algoritmom.

Zahvaljujem dr. sc. Danieli Kalafatović što me upoznala s biologijom i kemijom povezanom sa strukturom peptida i pružila mi dodatne izvore za edukaciju u tom području.

Zahvaljujem se prijateljima i obitelji za pomoć u svakodnevnim aktivnostima ali i onima na fakultetu.

Posebno se zahvaljujem sestri koja mi je dala ideju da se bavim računarstvom i koja mi je davala savjete o karijeri i tehničkim i društvenim aspektima rada na fakultetu.

Zahvaljujem se i svima ostalima koji su me podržavali, a ovdje ih nisam navela.

Sadržaj

Poglavlje 1

Uvod

Ovaj diplomski rad dio je uspostavljenog istraživačkog projekta Hrvatske zaklade za znanost pod naslovom "Dizajn katalitički aktivnih peptida i peptidnih nanostruktura" pod oznakom UIP-2019-04-7999. Rad u verziji na engleskom jeziku bit će poslan na razmatranje u prestižni časopis "Nature Machine Intelligence", a engleski tekst priložen je kao dodatak ovom radu.

Peptidi su molekule sastavljene od kombinacija 20 genski kodiranih aminokiselina i predstavljaju svestran alat za dobivanje supramolekularnih materijala s bogatom kemijskom i strukturnom raznolikošću [1]. Molekularno samosastavljanje (engl. Self-Assembly (SA)) temeljeno na slabim, nekovalentnim interakcijama čini jedan od temeljnih kemijskih procesa u živim organizmima, a naročito kada se događaju interakcije molekula peptida [2, 3]. Iako je ekspanzivan i dugotrajan, zbog eksponencijalnog rasta u prostoru pretraživanja za svaku aminokiselinu dodanu nizu, eksperimentalno otkriće novih samosastavljajućih peptida ostaje prevladavajući pristup [4].

Metode molekularne dinamike (engl. Molecular Dynamics (MD)) često su korištene za prevođenje laboratorijskih eksperimenata u *in silico* simulacije i smanjenje vremena potrebnog za ispitivanje jednog spoja [5, 6, 7, 8, 9], djelujući kao surogatni modeli za *in vitro* procjenu. Simulacije su korisne za stjecanje intuicije o rezultatima procjene, no konačna potvrda modeliranog procesa (npr. samosastavljanje peptida) zahtijeva laboratorijske testove. Iako je MD uspješno primijenjen na ispitivanje sklonosti agregaciji (engl. Aggregation Propensity (AP)) kao prekursora faze samosas-

tavljanja peptida za sve dipeptide [10] i tripeptide [1], proširenje ovog pristupa na tetrapeptide i šire i dalje je izazovno zbog dugačkog vremena izračuna [4, 11, 12].

Strojno učenje (engl. Machine Learning (ML)) pojavilo se kao učinkovita alternativa MD-u za *in silico* probir terapeutskih spojeva [13, 14, 15, 16, 17, 18, 19, 20, 21]. ML algoritmi brži su od MD simulacija i mogu postići prihvatljivu točnost s velikim skupovima podataka i pravim izborom arhitekture. Međutim, oni su uglavnom neistraženi za predviđanje sklonosti samosastavljanju peptida zbog nedostatka i neuravnoteženosti podataka [22].

Povratne neuronske mreže (engl. Recurrent Neural Network (RNN)) posebno su učinkovite u obradi sekvenci podataka zbog svoje sposobnosti obuhvaćanja sekvencijalnih ovisnosti, što ih razlikuje od drugih ML modela i mrežnih arhitektura [23]. Kao takvi, pronašli su široku primjenu u zaključivanju od sekvence do funkcije, kao što je analiza sentimenta u pisanom tekstu [24, 25], problem vrlo sličan predviđanju aktivnosti peptida. Stoga, RNN-ovi mogu pokazati performanse koje su superiorne u odnosu na tehnike ML-a kao što je metoda nasumične šume (engl. random forest), Support Vector Machine (SVM) i neuronske mreže koje nisu povratne za procjenu sklonosti SA modeliranjem sekvencijalnih odnosa s obzirom na kontekst između sastojaka peptida. U ovom radu uveden je pristup temeljen na RNN-u za procjenu samosastavljanja neklasificiranih peptida proizvoljne duljine koristeći samo peptidne sekvence kao ulazne podatke. Prednost ovog pristupa temelji se na upotrebi nejednoliko uzorkovanih značajki promjenjive duljine na temelju AP rezultata aminokiselina, dipeptida i tripeptida [26, 10, 1] kao značajki za bilo koji peptid od interesa. Ovaj model može nadopuniti ljudsku intuiciju u pokušaju identificiranja novih peptida s visokom sklonošću samosastavljanju na temelju nepristranog istraživanja prostora sekvenci potpomognutog ML-om.

U ovom radu trenirana je inačica modela koja prihvaća peptide s maksimalnom duljinom od 24 aminokiseline, kako bi odgovarala skupu podataka. Unatoč tome, model je svestran i može se primijeniti na nizove bilo koje duljine podešavanjem broja vrijednosti za ispunu tijekom treniranja. Pristupi temeljeni na strojnom učenju dosad su bili ograničeni skupovima podataka nedovoljne veličine i kvalitete [11]. Stoga je skup podataka koji obuhvaća peptide koji imaju sposobnost samosastavljanja i peptide koji nemaju sposobnost samosastavljanja sustavno prikupljen iz literature, što

je omogućilo treniranje modela na eksperimentalno potvrđenim sekvencama. Kombinacijom AP rezultata sa računalnim prikazom peptida temeljenom na sekvencijalnim svojstavima aminokiselina [27] (engl. Sequential Properties (SP)), izračunati su heterogeni podaci što je omogućilo formiranje većeg skupa podataka. Skup podataka korišten za treniranje modela neuronskih mreža sastojao se od 393 eksperimentalno potvrđena peptida, 249 onih koji imaju sposobnost samosastavljanja (SA) te 144 onih koji nemaju sposobnost samosastavljanja (engl. Non Self-Assembly (NSA)). Prisutna je neravnoteža u duljini sekvence jer većina peptida pripada skupini heksapeptida.

AP-ovi dipeptida i tripeptida čine ključne ulazne vrijednosti za modele generirane u ovom radu kojima je cilj predvidjeti sklonost samosastavljanju dužih sekvenci na temelju postojećeg znanja o minimalističkim sekvencama. U svrhu postizanja ovog cilja, trenirana su tri različita RNN modela koristeći različite značajke peptida na temelju: (i) AP rezultata, (ii) SP vrijednosti i (iii) hibridnog modela koji kombinira AP i SP. Prethodne studije [28] potvrdile su vrijednost korištenja AP rezultata za predviđanje samosastavljanja, stoga je upotrijebljen pristup kliznog prozora za izdvajanje AP vrijednosti aminokiselina, dipeptida i tripeptida koji su podskupovi izvorne sekvence. S druge strane, SP su se pokazali korisnim u obuhvaćanju i sekvencijalnih i fizikalno-kemijskih karakteristika sekvence te su se pokazali izrazito uspješnim načinom prikaza peptida za predviđanje njihove terapijske antimikrobne i antiviralne aktivnosti [27].

Kada se primjenjuju RNN-ovi na predviđanje svojstava peptida, važan korak prije treniranja modela sastoji se od predprocesiranja podataka, kako bi se dobili, odnosno postigli, strukturirani skupovi podataka skaliranjem, dodavanjem vrijednosti za ispunu do željene duljine itd. Posljedično, u koraku predprocesiranja, sve AP vrijednosti su skalirane na raspon $[-1, 1]$ i ispunjene do fiksne duljine niza od 24 vrijednošću 2, što je izvan raspona koji se koristi za podatke. Svim peptidima je nadodana ispuna do iste duljine radi obrade većeg broja instanci u istoj šarži (eng. batch), kako bi se povećala brzina i performanse neuronske mreže.

Tijekom treniranja korištena je unakrsna validacija u k preklopa (engl. k-fold cross-validation) kako bi se izbjegla pretjerana prilagodba (eng. overfitting) relativno maloj količini dostupnih podataka u fazi treniranja. Podjela je stratificirana

kako bi se osiguralo da preklopi, koji su izdvojeni iz podataka u kojima je prisutan neuravnotežen omjer SA i NSA klasa, sačuvaju isti postotak uzoraka za svaku klasu. Dodatno, ugniježđena unakrsna validacija u k preklopa korištena je za izvođenje optimizacije hiperparametara tijekom koje je testiran niz vrijednosti za višestruke hiperparametre, a samo oni najbolji odabrani su za konačni model.

Za sveobuhvatnu procjenu performansi neuronskih mreža korišten je niz metrika procjene koje obuhvaćaju različite aspekte ponašanja modela. Ove metrike uključivale su krivulje preciznosti i odziva (engl. Precision-Recall (PR)) i radne karakteristike prijamnika (engl. Receiver Operating Characteristic (ROC)), odgovarajuće područje ispod krivulja, točnost modela, rezultate F1 temeljene na vrijednostima preciznosti i odziva, kao i geometrijsku sredinu, stopu lažnih pozitivnih rezultata (engl. False Positive Rate (FPR)) i stopu stvarnih pozitivnih rezultata (engl. True Positive Rate (TPR)). ROC i PR pragovi ROC i PR određeni su tijekom procesa optimizacije hiperparametara na peptidima korištenim za validaciju u unutarnjoj petlji. Pragovi su definirani za pretvorbu izlaza modela, koji je SA vjerojatnost između 0 i 1, u binarnu klasu (0 ili 1, NSA ili SA).

Arhitektura duge kratkoročne memorije (engl. Long Short Term Memory (LSTM)) odabrana je jer se pokazala učinkovitom u rješavanju problema nestajanja gradijenta (engl. vanishing gradient problem), osiguravanju učinkovitog treniranja i poboljšanju performansi pri radu s opsežnim sekvencijalno ovisnim podacima [29]. Na temelju LSTM arhitekture razvijeno je pet modela na temelju AP vrijednosti, SP vrijednosti, hibridni model koji koristi i AP i SP vrijednosti, model koji koristi t-distribuirano stohastičko ugrađivanje susjeda (engl. t-Distributed Stochastic Neighbor Embedding (t-SNE)) na SP vrijednostima i model koji koristi t-SNE na AP i SP vrijednostima.

Podmodeli koji su uključivali AP vrijednosti kao ulaz primijenili su dvosmjerni LSTM sloj s 5 jedinica, a posljedična dimenzionalnost izlaznog tenzora bila je jednaka 10. Primijenjen je dodatni LSTM sloj koji nije dvosmjeran i ima 5 jedinica, s izlaznom tenzorskom dimenzionalnošću 5. Ovi podmodeli također su dodali gusto povezani sloj koji je imao 64, 96 ili 128 jedinica sa *selu* (skalirana eksponencijalna linearna jedinica) aktivacijom nakon ovoga tako da su izlazne dimenzije prije izostavljanja uzoraka iste u svim podmodelima. S druge strane, podmodeli koji su uključivali SP vrijednosti kao ulaz koristili su dva 1D konvolucijska sloja. Svaki konvolucijski sloj

stvorio je jezgru konvolucije koja je konvoluirana s ulazom sloja preko jedne prostorne (ili vremenske) dimenzije kako bi se proizveo izlazni tenzor. Postupak podešavanja hiperparametara postavljen je za testiranje veličine jezgre od 4, 6 ili 8. Svaki od dva konvolucijska sloja koristi 5 filtara, što znači da je dimenzionalnost izlaznog tenzora također 5. Taj je broj znatno manji od 64, 128 ili 256 filtara korištenih u modelu za klasifikaciju terapeutskih peptida [27]. Smanjenje složenosti modela bilo je potrebno zbog manjeg skupa podataka. Dvosmjerni LSTM sloj primijenjen u podmodelima na temelju SP vrijednosti imao je 32, 48 ili 64 jedinice, što znači da je dimenzionalnost izlaznog tenzora dvostruko veća i uključivala je 64, 96 ili 128 jedinica ovisno o odabranim hiperparametrima.

Kako bi se smanjila dimenzionalnost podataka kada se koristio SP, t-SNE [30] tehnika primijenjena je na 94 fizikalno-kemijska svojstva aminokiselina, dajući 3 meta-značajke. Obrazloženje za taj pristup bilo je svođenje broja značajki AP-a i SP-a na zajedničku ljestvicu, budući da postoje tri vrijednosti AP (aminokiseline, dipeptidi i tripeptidi) i 94 SP vrijednosti. Korištena je t-SNE tehnika s obzirom na njezinu utvrđenu upotrebu u istraživanju peptida za generiranje meta-značajki koje olakšavaju vizualizaciju podataka [31, 32]. Modeli neuronske mreže koji su koristili t-SNE tijekom predprocesiranja preferirali su veću veličinu jezgre (engl. kernel) (8 za SP model i 6 za hibridni AP-SP model) u LSTM slojevima u usporedbi s onima koji nisu (4 za SP model i hibridni AP-SP model). Slično tome, modeli neuronskih mreža koji su koristili t-SNE tijekom predprocesiranja preferirali su veći broj jedinica (64 za SP model i 48 za hibridni AP-SP model) u LSTM slojevima u usporedbi s onima koji nisu (32 za SP model i hibridni AP-SP model). Modeli koji imaju manji broj redaka u matrici koja predstavlja ulazne podatke manjak podataka kompenziraju većom dimenzionalnošću izlaznog prostora. Model koji kao ulaz koristi samo AP vrijednosti preferirao je najveći broj jedinica u svojim slojevima (128).

Model neuronske mreže koji je koristio samo SP vrijednosti ima najveću prosječnu točnost ($83,53\%$) i najmanji prosječni gubitak ($1,23 * 10^{-3}$) u svim epohama treniranja na temelju svih testiranih slučajnih početnih vrijednosti i svih testova tijekom optimizacije hiperparametara. Slijedi hibridni AP-SP model s prosječnom točnošću od $82,48\%$ i prosječnim gubitkom od $1,29 * 10^{-3}$. McNemarovim testom statističke značajnosti [33] i PR pragovima utvrđeno je da se rezultati klasifikacije

SP i hibridnih AP-SP modela ne razlikuju značajno ($P > 0,05$).

Provedeno je testiranje modela koristeći unakrsnu validaciju u k preklopa nakon čega su svi rezultati konkatenirani s ciljem sveobuhvatne i rigorozne analize performansi promatranih modela. Model neuronske mreže koji je koristio samo SP vrijednosti ima najveću površinu ispod ROC krivulje (engl. Area Under Curve (AUC)) ($0,861$), geometrijsku sredinu (engl. geometric mean (gmean)) ($0,788$), F1 rezultat ($0,848$) i točnost (80,1%) na temelju ROC praga za samosastavljanje (0.617). Odmah iza njega slijedi hibridni AP-SP model s PR AUC rezultatom od $0,857$, geometrijskom sredinom od $0,776$, F1 rezultatom od $0,841$ i točnošću od $79,1\%$ (ROC prag za AP-SP model iznosi 0.607).

Model neuronske mreže koji je koristio samo SP vrijednosti također ima najveću geometrijsku sredinu ($0,704$), F1 rezultat ($0,860$) i točnost ($79,7\%$) na temelju PR praga za samosastavljanje (0.321). Ponovo ga tijesno prati hibridni AP-SP model s geometrijskom sredinom od $0,692$, F1 rezultatom od $0,850$ i točnošću od $78,3\%$ (PR prag za AP-SP model iznosi 0.320). Hibridni AP-SP model ima malo veći AUC ($0,924$) od SP modela ($0,917$) na temelju PR krivulje.

Nadalje, model neuronske mreže koji je koristio samo SP vrijednosti ima najveću geometrijsku sredinu ($0,766$), F1 rezultat ($0,856$) i točnost ($80,4\%$) na temelju praga za samosastavljanje od 0,5. Konačno ga tijesno slijedi hibridni AP-SP model s geometrijskom sredinom od $0,758$, rezultatom F1 od $0,849$ i točnošću od $79,5\%$. Prag od 0,5 korišten je kao standardna vrijednost bez podešavanja budući da dijeli mogući raspon vjerojatnosti samosastavljanja između 0 i 1 točno na pola.

Provedeno je dodatno testiranje modela u potpuno novom okruženju izvedenom iz podataka nedavne studije koja je pokazala prednosti upotrebe umjetne inteligencije za otkrivanje samosastavljajućih peptida [4]. U toj studiji dizajnirano je 20 heksapeptida za koje su dobiveni izračuni AP rezultata na temelju MD-a zajedno s eksperimentalno potvrđenim statusom agregacije (1 ili 0). Kako bi se razdvojilo istinski pozitivnu (engl. True Positive (TP)) i istinski negativnu (engl. True Negative (TN)) klasu s minimalnim brojem lažno pozitivnih (engl. False Positive (FP)) i lažno negativnih (engl. False Negative (FN)), procijenjena je granična vrijednost AP ocjene od $1,765$, što također odgovara prosječnom AP-u heksapeptida za koje je provedena opsežna analiza primjenom tehnike molekularne dinamike. Granična

vrijednost omogućila je procjenu performansi unaprijed treniranih modela u zadatku binarne klasifikacije za novi skup podataka od 6578 heksapeptida bez eksperimentalno ispitanog statusa agregacije.

AP model dominantan je prediktor kada se promatraju AUC rezultati ($ROC\ AUC = 0,835$, $PR\ AUC = 0,862$) i koeficijenti korelacije ($Pearson = 0,60$ i $Spearman = 0,66$). Ovaj je ishod očekivan budući da je ovaj model treniran na AP vrijednostima aminokiselina, dipeptida i tripeptida, a izlaz predviđanja povezan je s AP vrijednostima nešto duljih sekvenci heksapeptida. Razina korelacije može se protumačiti kao umjerena, otkrivajući da bolje predviđanje zahtijeva više od samo AP vrijednosti podskupova sastavne sekvence. Kada se uspoređuje izbor pragova za konačnu klasifikaciju, AP model je u skladu s prethodnom analizom i daje najbolje performanse korištenjem PR praga ($gmean = 0,743$, $F1 = 0,803$ i $Acc = 76,3\%$). Nasuprot tome, t-SNE AP-SP model ima najbolje performanse s ROC pragom ($gmean = 0,701$, $F1 = 0,683$ i $Acc = 69,5\%$), zajedno sa SP modelom ($F1 = 0,684$). Smanjenje broja SP vrijednosti pomoću t-SNE tehnike omogućilo je modelu da dobije druge najbolje razine korelacije ($Pearson = 0,56$, $Spearman = 0,59$) i AUC vrijednosti ($ROC\ AUC = 0,808$, $PR\ AUC = 0,834$) nakon AP modela te najbolje rezultate sa standardnim pragom ($gmean = 0,715$, $F1 = 0,710$ i $Acc = 70,8\%$), odmah nakon SP modela ($F1 = 0,708$).

Stručnjaci su predvidjeli sposobnost samosastavljanja za 11 peptida, a umjetna inteligencija (engl. Artificial Intelligence (AI)) je to predvidjela za dodatnih 9 sekvenci korištenjem modela autora Batra et al. [4], dovodeći ukupan broj sekvenci heksapeptida koje je opsežno proučavao MD na 20. Modeli razvijeni u ovom radu predviđali su vjerojatnost samosastavljanja za oba skupa sekvenci. Ljudsko predviđanje samosastavljanja za 11 sekvenci imalo je točnost od 55%. Svi modeli razvijeni u ovom radu bili su uspješniji s obzirom na točnost na 20 korištenih sekvenci te je PR prag za svaki model primijenjen za pretvaranje predviđanja u binarne klase. S druge strane, modeli razvijeni u ovom radu, osim t-SNE SP modela, koji ima točnost od 60% na 20 sekvenci, pokazali su superiornu točnost u odnosu na model autora Batra et al., koji je na 9 heksapeptida imao točnost od 67%. Budući da su ljudska predviđanja i predviđanja modela autora Batra et al. predviđala samo pozitivnu klasu, stopa istinski negativnih rezultata (engl. True Negative Rate (TNR))

i stopa lažno negativnih rezultata (engl. False Negative Rate (FNR)) su jednake 0, geometrijska sredina je također jednaka 0, dok je točnost jednaka TPR-u . Iz ovih vrijednosti procijenjene su geometrijska sredina i F1 rezultati te su uspoređeni s modelima razvijenima u ovom radu. Svi modeli razvijeni u ovom radu nadmašuju ljudska predviđanja na temelju njihovog F1 rezultata. SP model nadmašuje predviđanja modela autora Batra et al. na temelju F1 rezultata. Svi modeli razvijeni u ovom radu imaju iste ili bolje rezultate od ljudskih predviđanja i predviđanja modela Batra et al. na temelju njihove geometrijske sredine.

Bibliografija

[1] P. W. Frederix, G. G. Scott, Y. M. Abul-Haija, D. Kalafatovic, C. G. Pappas, N. Javid, N. T. Hunt, R. V. Ulijn, and T. Tuttle, "Exploring the sequence space for (tri-) peptide self-assembly to design and discover new hydrogels," *Nature chemistry*, vol. 7, no. 1, pp. 30–37, 2015.

[2] A. Lampel, "Biology-inspired supramolecular peptide systems," *Chem*, vol. 6, no. 6, pp. 1222–1236, 2020.

[3] P. Janković, I. Šantek, A. S. Pina, and D. Kalafatovic, "Exploiting peptide self-assembly for the development of minimalistic viral mimetics," *Frontiers in Chemistry*, vol. 9, p. 723473, 2021.

[4] R. Batra, T. D. Loeffler, H. Chan, S. Srinivasan, H. Cui, I. V. Korendovych, V. Nanda, L. C. Palmer, L. A. Solomon, H. C. Fry *et al.*, "Machine learning overcomes human bias in the discovery of self-assembling peptides," *Nature chemistry*, pp. 1–9, 2022.

[5] G. Gocheva, K. Peneva, and A. Ivanova, "Self-assembly of doxorubicin and a drug-binding peptide studied by molecular dynamics," *Chemical Physics*, vol. 525, p. 110380, 2019.

[6] C. Guo, Y. Luo, R. Zhou, and G. Wei, "Triphenylalanine peptides self-assemble into nanospheres and nanorods that are different from the nanovesicles and nanotubes formed by diphenylalanine peptides," *Nanoscale*, vol. 6, no. 5, pp. 2800–2811, 2014.

[7] O.-S. Lee, V. Cho, and G. C. Schatz, "Modeling the self-assembly of peptide amphiphiles into fibers using coarse-grained molecular dynamics," *Nano letters*, vol. 12, no. 9, pp. 4907–4913, 2012.

[8] C. A. Hauser, R. Deng, A. Mishra, Y. Loo, U. Khoe, F. Zhuang, D. W. Cheong, A. Accardo, M. B. Sullivan, C. Riekel *et al.*, "Natural tri-to hexapeptides self-assemble in water to amyloid β-type fiber aggregates by unexpected α-helical

intermediate structures," *Proceedings of the National Academy of Sciences*, vol. 108, no. 4, pp. 1361–1366, 2011.

[9] P. W. Frederix, I. Patmanidis, and S. J. Marrink, "Molecular simulations of self-assembling bio-inspired supramolecular systems and their connection to experiments," *Chemical Society Reviews*, vol. 47, no. 10, pp. 3470–3489, 2018.

[10] P. W. Frederix, R. V. Ulijn, N. T. Hunt, and T. Tuttle, "Virtual screening for dipeptide aggregation: Toward predictive tools for peptide self-assembly," *The journal of physical chemistry letters*, vol. 2, no. 19, pp. 2380–2384, 2011.

[11] P. Zhou, C. Yuan, and X. Yan, "Computational approaches for understanding and predicting the self-assembled peptide hydrogels," *Current Opinion in Colloid & Interface Science*, p. 101645, 2022.

[12] N. Palmer, J. R. Maasch, M. D. Torres, and C. de la Fuente-Nunez, "Molecular dynamics for antimicrobial peptide discovery," *Infection and Immunity*, vol. 89, no. 4, pp. e00 703–20, 2021.

[13] W.-F. Zeng, X.-X. Zhou, S. Willems, C. Ammar, M. Wahle, I. Bludau, E. Voytik, M. T. Strauss, and M. Mann, "Alphapeptdeep: a modular deep learning framework to predict peptide properties for proteomics," *Nature Communications*, vol. 13, no. 1, pp. 1–14, 2022.

[14] S. N. H. Bukhari, J. Webber, and A. Mehbodniya, "Decision tree based ensemble machine learning model for the prediction of zika virus t-cell epitopes as potential vaccine candidates," *Scientific Reports*, vol. 12, no. 1, pp. 1–11, 2022.

[15] M. C. Melo, J. R. Maasch, and C. de la Fuente-Nunez, "Accelerating antibiotic discovery through artificial intelligence," *Communications biology*, vol. 4, no. 1, p. 1050, 2021.

[16] J. Chen, H. H. Cheong, and S. W. Siu, "Xdeep-acpep: deep learning method for anticancer peptide activity prediction based on convolutional neural network and multitask learning," *Journal of chemical information and modeling*, vol. 61, no. 8, pp. 3789–3803, 2021.

[17] S. Akbar, A. Ahmad, M. Hayat, A. U. Rehman, S. Khan, and F. Ali, "iatbp-hyb-enc: Prediction of antitubercular peptides via heterogeneous feature representation and genetic algorithm based ensemble learning model," *Computers in Biology and Medicine*, vol. 137, p. 104778, 2021.

[18] P. G. Aronica, L. M. Reid, N. Desai, J. Li, S. J. Fox, S. Yadahalli, J. W. Essex, and C. S. Verma, "Computational methods and tools in antimicrobial peptide

research," *Journal of Chemical Information and Modeling*, vol. 61, no. 7, pp. 3172–3196, 2021.

[19] M. M. Hasan, N. Schaduangrat, S. Basith, G. Lee, W. Shoombuatong, and B. Manavalan, "Hlppred-fuse: improved and robust prediction of hemolytic peptide and its activity by fusing multiple feature representation," *Bioinformatics*, vol. 36, no. 11, pp. 3350–3356, 2020.

[20] B. Manavalan, T. H. Shin, M. O. Kim, and G. Lee, "Aippred: sequence-based prediction of anti-inflammatory peptides using random forest," *Frontiers in pharmacology*, vol. 9, p. 276, 2018.

[21] M. Attique, M. S. Farooq, A. Khelifi, and A. Abid, "Prediction of therapeutic peptides using machine learning: computational models, datasets, and feature encodings," *IEEE Access*, vol. 8, pp. 148 570–148 594, 2020.

[22] F. Li, J. Han, T. Cao, W. Lam, B. Fan, W. Tang, S. Chen, K. L. Fok, and L. Li, "Design of self-assembly dipeptide hydrogels and machine learning via their chemical features," *Proceedings of the National Academy of Sciences*, vol. 116, no. 23, pp. 11 259–11 264, 2019.

[23] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE transactions on neural networks*, vol. 5, no. 2, pp. 157–166, 1994.

[24] G. Yang, Y. Jiayu, X. Dongdong, G. Zelin, and H. Hai, "Feature-enhanced text-inception model for chinese long text classification," *Scientific Reports*, vol. 13, no. 1, p. 2087, 2023.

[25] L. Zhang, S. Wang, and B. Liu, "Deep learning for sentiment analysis: A survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 8, no. 4, p. e1253, 2018.

[26] N. De Groot, I. Pallarès, F. Avilés, J. Vendrell, and S. Ventura, "Prediction of hot spots of aggregation in disease-linked polypeptides„" *BMC structural biology*, vol. 5, p. 18, 02 2005.

[27] E. Otović, M. Njirjak, D. Kalafatovic, and G. Mauša, "Sequential properties representation scheme for recurrent neural network-based prediction of therapeutic peptides," *Journal of Chemical Information and Modeling*, vol. 62, no. 12, pp. 2961–2972, 2022.

[28] O. Conchillo-Solé, N. De Groot, F. Avilés, J. Vendrell, X. Daura, and S. Ventura, "AGGRESCAN: a server for the prediction of "hot spots" of aggregation in polypeptides," *BMC bioinformatics*, vol. 8, p. 65, 02 2007.

[29] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," in *International conference on machine learning*. Pmlr, 2013, pp. 1310–1318.

[30] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.

[31] L. Wei, X. Ye, T. Sakurai, Z. Mu, and L. Wei, "Toxibtl: prediction of peptide toxicity based on information bottleneck and transfer learning," *Bioinformatics*, vol. 38, no. 6, pp. 1514–1524, 2022.

[32] S. N. Dean, J. A. E. Alvarez, D. Zabetakis, S. A. Walper, and A. P. Malanoski, "Pepvae: variational autoencoder framework for antimicrobial peptide generation and activity prediction," *Frontiers in Microbiology*, vol. 12, p. 725727, 2021.

[33] Q. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages," *Psychometrika*, vol. 12, no. 2, pp. 153–157, 1947.

Pojmovnik

AI Artificial Intelligence. 7

AP Aggregation Propensity. 1–7

AUC Area Under Curve. 6, 7

FN False Negative. 6

FNR False Negative Rate. 8

FP False Positive. 6

FPR False Positive Rate. 4

gmean geometric mean. 6, 7

LSTM Long Short Term Memory. 4, 5

MD Molecular Dynamics. 1, 2, 6, 7

ML Machine Learning. 2

NSA Non Self-Assembly. 3, 4

PR Precision-Recall. 4–7

RNN Recurrent Neural Network. 2, 3

ROC Receiver Operating Characteristic. 4, 6, 7

SA Self-Assembly. 1–4

SP Sequential Properties. 3–7

SVM Support Vector Machine. 2

t-SNE t-Distributed Stochastic Neighbor Embedding. 4, 5, 7

TN True Negative. 6

TNR True Negative Rate. 7

TP True Positive. 6

TPR True Positive Rate. 4, 8

Sažetak

Predložen je pristup za rješavanje problema heterogenih podataka iz sekvencijalnih svojstava (SP) i ocjene sklonosti agregaciji (AP) za peptide sastavljene od 1, 2 i 3 aminokiseline uporabom povratnih neuronskih mreža koje obrađuju sekvence promjenjive duljine koristeći svoju unutarnju memoriju. Kako bi se postigao jednak broj AP i SP vrijednosti, t-distribuirano stohastičko ugrađivanje susjeda (t-SNE) primijenjeno je na 94 SP vrijednosti i izdvojene su 3 meta-značajke. Pragovi temeljeni na krivuljama ROC (radna karakteristika prijamnika) i PR (preciznost-odziv) testirani su kako bi se unaprijedila binarna klasifikacija. Visoki rezultati geometrijske sredine ($0,704$ za SP model i PR prag) koji su vrlo blizu rezultata točnosti ($79,7\%$ za SP model i PR prag) dokazuju da je predviđanje moguće za pozitivnu i negativnu klasu samosastavljajućih peptida. Neklasificirani heksapeptidi korišteni su za testiranje primjenjivosti modela za predviđanje ocjene AP. Najuspješniji među razvijenim modelima, SP model, nadmašuje ljudska predviđanja i predviđanja najsuvremenijeg modela iz literature na temelju točnosti, F1 rezultata i geometrijske sredine. Modeli bi se mogli koristiti kao dopuna ljudskoj intuiciji u stvaranju novih peptidnih sekvenci s velikom sklonošću samosastavljanju.

Ključne riječi — **Peptidi, Neuronske mreže, Sekvencijalna svojstva, Sklonost agregaciji, Predviđanje samosastavljanja**

Abstract

We propose an approach to tackle the issue of heterogeneous data from sequential properties (SP) and aggregation-propensity (AP) scores for peptides composed of 1, 2, and 3 amino acids by using Recurrent Neural Networks that process sequences of variable length using their internal memory. To achieve an equal number of AP and SP values, t-distributed stochastic neighbour embedding (t-SNE) was applied to the 94 SP values, and 3 meta-features were extracted. Thresholds based on ROC (Receiver Operating Characteristic) and PR (Precision-Recall) curves were tested to advance binary classification. High geometric mean scores (0.704 for the SP model

and PR threshold) that are very close to accuracy scores (79.7% for the SP model and PR threshold) prove that prediction is possible both for the positive and negative class of peptides based on self-assembly. Unclassified hexapeptides were used to test whether the model was applicable for predicting AP scores. The most successful of the developed models, the SP model, outperforms human predictions and the predictions of the state-of-the-art model from the literature based on accuracy, F1 score, and geometric mean. The models could complement human intuition in generating novel peptide sequences with a high propensity to self-assemble.

Keywords **— Peptides, Neural networks, Sequential properties, Aggregation propensity, Self-assembly prediction**

Dodatak A

Prediction of peptide-based supramolecular systems: Building a Neural Network-based model using heterogeneous data

Prediction of peptide-based supramolecular systems: Building a Neural Network-based model using heterogeneous data

Lucija Žužić[1], Marko Njirjak[1], Daniela Kalafatovic[2,3*] and Goran Mauša[1*]

[1]University of Rijeka, Faculty of Engineering, Vukovarska 58, Rijeka, 51000, Croatia.
[2]University of Rijeka, Department of Biotechnology, R. Matejcic 2, Rijeka, 51000, Croatia.
[3]University of Rijeka, Center for Artificial Intelligence and Cybersecurity, R. Matejcic 2, Rijeka, 51000, Croatia.

*Corresponding author(s). E-mail(s): daniela.kalafatovic@uniri.hr; gmausa@riteh.hr;
Contributing authors: lzuzic@riteh.hr; mnjirjak@riteh.hr;

Abstract

We propose an approach to tackle the issue of heterogeneous data from sequential properties (SP) and aggregation-propensity (AP) scores for peptides composed of 1, 2, and 3 amino acids by using Recurrent Neural Networks that process sequences of variable length using their internal memory. To achieve an equal number of AP and SP values, t-distributed stochastic neighbour embedding (t-SNE) was applied to the 94 SP values, and 3 meta-features were extracted. Thresholds based on ROC (Receiver Operating Characteristic) and PR (Precision-Recall) curves were tested to advance binary classification. High geometric mean scores (0.704 for the SP model and PR threshold) that are very close to accuracy scores (79.7% for the SP model and PR threshold) prove that prediction is possible both for the positive and negative class of peptides based on self-assembly. Unclassified hexapeptides were used to test whether the model was applicable for predicting AP scores. The most successful of the developed models, the SP model, outperforms human predictions and the predictions of the state-of-the-art model from the literature based on accuracy, F1 score, and geometric

mean. The models could complement human intuition in generating novel peptide sequences with a high propensity to self-assemble.

Keywords: peptides, neural networks, sequential properties, aggregation propensity, self-assembly prediction

1 Main

Molecular self-assembly (SA) based on weak, non-covalent, interactions constitutes one of the fundamental chemical processes in living organisms [1, 2]. Peptides are molecules composed of combinations of 20 gene-encoded amino acids and constitute a versatile toolbox for obtaining supramolecular materials with rich chemical and structural diversities [3]. It is no surprise that the organization of peptidic building blocks into three-dimensional structures results in materials with tremendous complexities and emerging properties such as catalysis and molecular recognition [4–6]. Consequently, the rational and computational design of peptide-based materials accompanied by extensive experimental validations established relevant and applicable design rules leading to great progress in the obtaining of self-assembling materials used for a wide range of applications [3, 7–9]. However, the exact sequence-structure-to-function relationships still remain beyond our comprehension.

Although expansive and time-consuming, due to an exponential growth in the search space for each amino acid added to the sequence, the experimental discovery of new self-assembling peptides remains the predominant approach [10]. However, this NP-hard discovery process advocates alternative methods for efficient navigation of such vast chemical spaces [11]. Furthermore, laboratory investigations of supramolecular peptide nanostructures employing electron or atomic force microscopy, including synthesis, purification, and sample preparation, require highly trained experts and sophisticated instrumentation [12–17] and it can take months to obtain a fully characterized peptide. On the other hand, the intractability of exhaustively examining the entire search space when considering peptides longer than three amino acids, and the sparseness of useful molecules in such spaces [18] yielded design rules for peptide self-assembly, namely patterning strategies manipulating hydrophobic-hydrophilic balance [19] and molecular templating [20]. Such procedures aim to constrain the peptide design and reduce the number of peptides that reside in the available search space to a more manageable level. Nevertheless, they introduce an unwanted bias towards specific regions of the search space, thereby disregarding potentially promising areas and limiting the discovery of novel sequences.

Molecular dynamics (MD) methods have been extensively used to translate laboratory experiments into *in silico* simulations and reduce the time required for a single compound examination [21–25], acting as surrogate models for *in vitro* evaluation. An inevitable consequence is the introduction of

errors into the results, the degree of which partially depends on the simulation setup details (e.g. coarse-grained MD [26], simulation time, and system size [27]). Therefore, MD simulations are advantageous for gaining intuition about evaluation results, yet the definitive confirmation of the modeled process (e.g. peptide self-assembly) requires experimental validation. Although MD has been successfully applied to the examination of aggregation propensity (AP) as a precursor stage of peptide self-assembly for all dipeptides [27] and tripeptides [3], extending this approach to tetrapeptides and beyond remains challenging due to high computational costs [10, 28, 29].

Machine learning (ML) has emerged as an efficient alternative to MD for the *in silico* screening of therapeutic compounds [30–38]. ML algorithms run faster than MD simulations and can achieve acceptable accuracy with large datasets and the right choice of architecture. However, they are mostly unexplored for the prediction of peptide self-assembly propensity due to the scarcity and imbalance of data [39]. In recent studies, ML-based sequence preselections based on support vector machine (SVM) and decision tree models were applied to guide the search for the most promising peptides, demonstrating its advantages over human experts and mitigating design biases they are prone to [10]. This approach yielded higher average AP scores compared to the exhaustive search, and their scalability made them applicable to search spaces composed of lengthy sequences, including octapeptides and proteins [40]. Although these studies employed ML, they relied on MD for the final AP score calculations. As we advance towards exploring molecular search spaces of increasing size, MD simulations, and their setup can become a bottleneck and, therefore, it is favorable to minimize their usage for unbiased and fast searches of the peptide space [41, 42].

Recurrent neural networks (RNNs) are particularly effective in processing data sequences due to their capacity to capture sequential dependencies, which distinguishes them from other ML models and network architectures [43]. As such, they found a widespread application in sequence-to-function inference, such as sentiment analysis [44, 45], a problem highly similar to the prediction of peptide activity. Therefore, RNNs can exhibit performance superior to those of ML techniques such as random forest, SVM, and non-recurrent neural networks for SA propensity assessment by modeling context-aware sequential relationships between peptide constituents. In this paper, we introduce an RNN-based approach for self-assembly assessment of unclassified peptides of arbitrary length employing only peptide sequences as input. The advantage of this approach is based on the use of irregularly sampled features of unequal length based on the AP scores of amino acids, dipeptides, and tripeptides [3, 27, 46] as predictor variables for any peptide of interest.

2 Results and Discussion

The prediction of the self-assembly propensity of arbitrary-length peptides constitutes a supervised learning problem. One approach to address it is

through the development of RNN prediction models where the SA probability is predicted from the peptide sequence (Figure 1). ML-based approaches were thus far limited by datasets of insufficient size and quality [28]. Hence, a dataset encompassing self-assembling and non-assembling peptides was manually curated from the literature, allowing the model to be trained on experimentally validated sequences. By combining the AP scores with the sequential properties representation scheme [47] (SP), heterogeneous data were computed allowing a larger dataset to be formed. The dataset used for the training of neural network models consisted of 393 experimentally validated peptides, labeled SA (self-assembling; 249) and NSA (non-assembling; 144). A considerable imbalance in sequence length is evident, with most of the peptides belonging to the hexapeptide group (Figure 2.a).

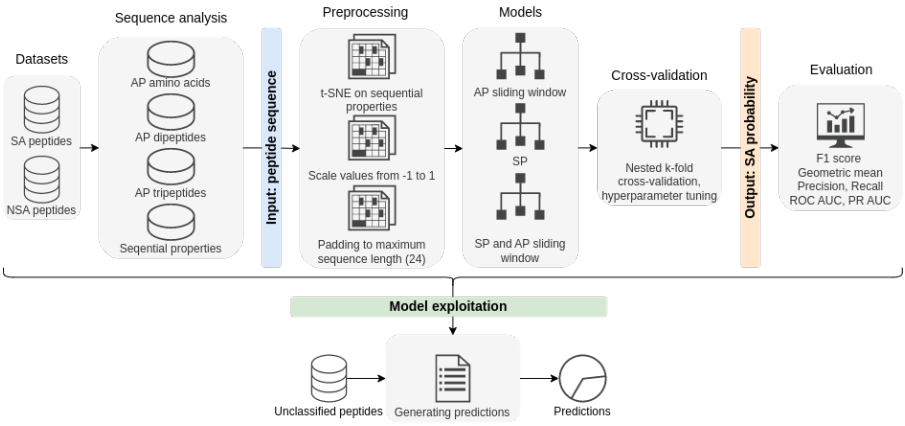


Fig. 1 Overview of the proposed prediction model. The model was based on heterogeneous data obtained by applying sliding windows of varying length: from single amino acids to dipeptides and tripeptides. Three models are trained on SA and NSA peptides expressed through AP and SP values. Data pre-processing based on the t-SNE technique was performed for dimensionality reduction, scaling the values to match the [-1, 1] range for better performance, and padding shorter sequences to expedite training. Models are optimized and thoroughly evaluated using the nested k-fold cross-validation sampling technique before yielding a final model ready for exploitation.

Amino acids such as phenylalanine, di- and tri- peptides including FF, YY, and WW and tripeptides such as FFF, CFF, PFF FFV, VFF, LFF, KYF, KYY are the simplest building blocks used in peptide nanotechnology to obtain highly organized supramolecular materials with varying architectures, with diphenylalanine (FF) being the most widely studied [2]. FF and VFF are key motifs identified in β-amyloids, present in Alzheimer's disease [48]. The amino acid composition, the physico-chemical properties, and the position of a particular amino acid within the sequence and its neighbouring residues are some of the factors that affect the formation of supramolecular assemblies and their morphology [2, 49]. Although distinct sequence patterns exist, it is

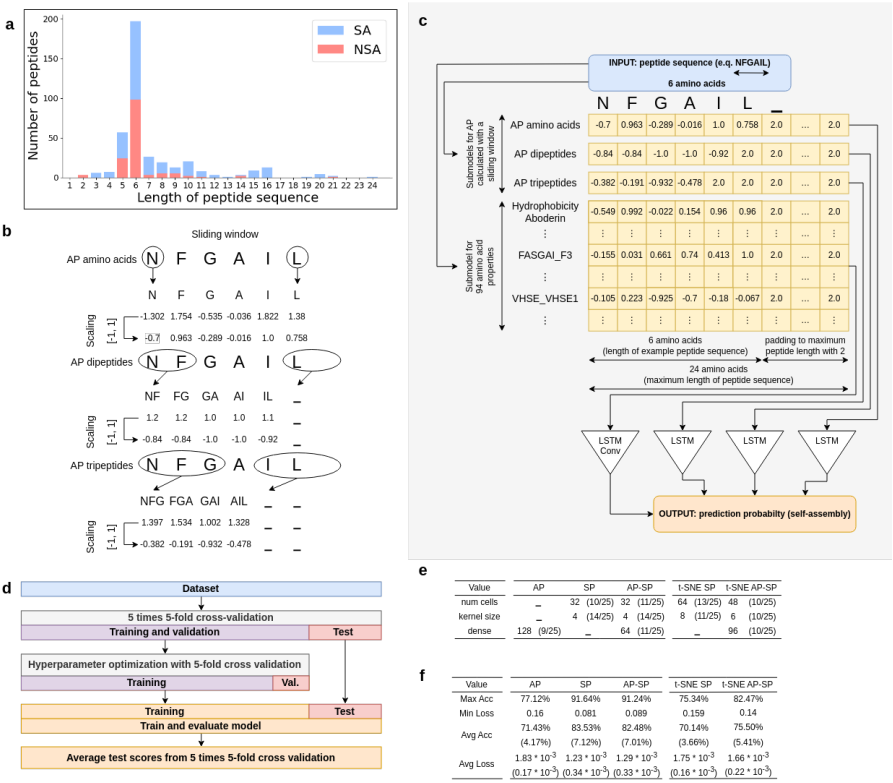
still challenging to attribute the precise self-assembly propensity based on the amino acid sequence. For this reason, di- and tri- peptide APs constitute key input values for the models generated in this study that aim to predict the self-assembly propensity of longer sequences based on the existing knowledge of minimalistic ones.

In pursuit of this goal, three distinct RNN models were trained by leveraging different peptide features based on: (i) AP scores, (ii) SP values, and (iii) a hybrid model combining both AP and SP. Previous studies [48] have demonstrated the value of using AP scores as SA predictors, therefore, we used a sliding window approach to extract the AP scores of contiguous amino acids, di-, and tri- peptides within the original sequence. SP proved useful in capturing both the sequence's sequential and physico-chemical characteristics. SP values were calculated for each peptide sequence, as previously applied by our group to therapeutic peptides [47]. The currently trained version of the model accepts peptides with a maximum length of 24 amino acids, to match the dataset. Nevertheless, the model is versatile and can be applied to sequences of any length by adjusting the padding value during training.

2.1 Building the prediction model and parameters fine-tuning

RNNs are deep learning algorithms that are often used for sequential or temporal problems, such as language translation and processing. They are applied in applications including Apple Siri [50] and Google Translate [51]. Their ability to memorize previously received inputs due to their internal memory constitutes an important advantage meaning that their outputs are influenced by prior inputs. When applying RNNs to predicting peptide properties, an important step prior to model training consists of data preprocessing, to obtain/achieve structured datasets through scaling, padding, etc. Consequently, in the preprocessing step, all AP values were scaled to a range $[-1, 1]$ and padded to a fixed sequence length of 24 with a value of 2, which is outside the range for the data. All peptides were padded to the same length to allow processing in the same batch, consequently increasing the speed and performance of the neural network. The sliding window was used to identify minimal sequence building motifs for which AP and SP values were computed (as shown in Figure 2.b). An example of the input data for a hybrid AP-SP model is included in Figure 2.c, with 94 SP and three AP values (amino acids, dipeptides, and tripeptides).

During training, stratified k-fold cross-validation was used to avoid overfitting that might be caused by the small dataset size and to ensure that folds, which are split from the imbalanced SA-NSA ratio, preserve the same percentage of samples for each class. Additionally, nested k-fold cross-validation was used to perform hyperparameter optimization during which a number of values for multiple hyperparameters were tested and only the best were selected for the final model, as described in Figure 2.d. The preferred values for each hyperparameter in each model are shown in Figure 2.e. For a comprehensive



Value	AP	SP	AP-SP	t-SNE SP	t-SNE AP-SP
num cells	_	32 (10/25)	32 (11/25)	64 (13/25)	48 (10/25)
kernel size	_	4 (14/25)	4 (14/25)	8 (11/25)	6 (10/25)
dense	128 (9/25)	_	64 (11/25)	_	96 (10/25)

Value	AP	SP	AP-SP	t-SNE SP	t-SNE AP-SP
Max Acc	77.12%	91.64%	91.24%	75.34%	82.47%
Min Loss	0.16	0.081	0.089	0.159	0.14
Avg Acc	71.43% (4.17%)	83.53% (7.12%)	82.48% (7.01%)	70.14% (3.66%)	75.50% (5.41%)
Avg Loss	$1.83 * 10^{-3}$ ($0.17 * 10^{-3}$)	$1.23 * 10^{-3}$ ($0.34 * 10^{-3}$)	$1.29 * 10^{-3}$ ($0.33 * 10^{-3}$)	$1.75 * 10^{-3}$ ($0.16 * 10^{-3}$)	$1.66 * 10^{-3}$ ($0.22 * 10^{-3}$)

Fig. 2 Neural network data pre-processing and hyperparameter optimization. **a** Input peptide sequences by length and experimentally confirmed self-assembly status. **b** Schematic representation of the pre-processing of peptide sequences with a sliding window using amino acids, dipeptides, and tripeptides. **c** Structure of the input data for the example sequence (NFGAIL) for a hybrid AP-SP neural network model. **d** Diagram of the workflow during the construction of neural network models. **e** Most commonly selected values during hyperparameter optimization along with the number of occurrences. **f** Maximum accuracy, minimal loss, and average accuracy and loss during training, along with standard deviation.

assessment of the neural networks' performance, a range of evaluation metrics was utilized that capture various aspects of the models' behavior. These metrics included precision-recall (PR) and receiver operating characteristic (ROC) curves, the corresponding area under the curves, model accuracy, F1 scores based on precision and recall values, as well as the geometric mean of false positive rate (FPR) and true positive rate (TPR).

During hyperparameter selection, training loss, validation loss, training accuracy, and validation accuracy are plotted for each model. The minimal and average training loss, as well as the maximal and average training accuracy, are recorded in Figure 2.f. The model that shows the largest training accuracy and training loss can be expected to perform the best during testing, which was proven to be true for the SP model. The neural network model that used only SP values has the highest average accuracy (83.53%) and the lowest average

loss ($1.23 * 10^{-3}$) in all training epochs based on all random seeds tested and all tests during hyperparameter optimization. It is closely followed by the hybrid AP-SP model with an average accuracy of 82.48% and an average loss of $1.29 * 10^{-3}$. The classification results of the SP and hybrid AP-SP models were found not to be significantly different ($P > 0.05$) using McNemar's statistical significance test [52] and PR thresholds. For models that are trained after hyperparameters are chosen, only training loss and accuracy are plotted, as the validation set of peptides is no longer needed once the hyperparameters are chosen.

The LSTM architecture was chosen because it proved effective in addressing the vanishing gradient problem, ensuring efficient training, and improving performance when dealing with extensive sequentially-dependent data [53]. Based on the LSTM architecture, five models were developed based on AP values, SP values, a hybrid model that uses both AP and SP values, a model that uses t-SNE on SP values, and a model that uses t-SNE on hybrid AP and SP.

The submodels that included AP values as input applied a bidirectional LSTM layer with 5 units, and the consequent dimensionality of the output tensor was equal to 10. An additional LSTM layer that is not bidirectional and has 5 units was applied, with an output tensor dimensionality of 5. These submodels also added a densely connected layer that had 64, 96, or 128 units with selu (scaled exponential linear unit) activation after this so that the output dimensions before dropout are the same in all submodels. On the other hand, the submodels that included SP values as input used two 1D convolutional layers. Each convolutional layer created a convolution kernel that is convolved with the layer input over a single spatial (or temporal) dimension to produce a tensor of outputs. The hyperparameter tuning procedure was set up to test a kernel size of 4, 6, or 8. Each of the two convolutional layers uses 5 filters, which means that the dimensionality of the output tensor is also 5. This number is significantly lower than the 64, 128, or 256 filters used in the model for classifying therapeutic peptides [47]. The reduction in model complexity was required due to the smaller data set. The bidirectional LSTM layer applied in the submodels based on SP values had 32, 48, or 64 units, which means that the dimensionality of the output tensor was twice as large and included 64, 96, or 128 units depending on the selected hyperparameters. All layers of the different RNN model architectures are visible in Figure 3.

To reduce the dimensionality of the data when SP was used, t-distributed stochastic neighbour embedding (t-SNE) [54] was applied to the 94 physicochemical amino acid properties, yielding 3 meta-features. The rationale behind the approach was to bring the number of AP and SP features to a common scale, by applying t-SNE, given its established use in peptide research to generate meta-features and facilitate data visualization [55, 56]. The neural network models that used t-SNE during preprocessing preferred a larger kernel size (8 for the SP model and 6 for the hybrid AP-SP model) in the LSTM layers compared to those that did not (4 for both the SP model and the hybrid

AP-SP model). Similarly, neural network models that used t-SNE during preprocessing preferred a larger number of units (64 for the SP model and 48 for the hybrid AP-SP model) in the LSTM layers compared to those that did not (32 for both the SP model and the hybrid AP-SP model). Models that have a smaller number of rows in the matrix that represents the input data compensate for the lack of data with a larger dimensionality of the output space. According to this claim, the model that uses only AP values as input preferred the largest number of units in its layers (128) compared to all other models.

2.2 Testing the models: SP outperforms AP

Histograms depicting the predicted self-assembly probability for tested peptides with and without self-assembly are generated for the models that are trained after choosing the hyperparameters, as shown in Figure 3. The histograms depicting the SA probability show that the AP model was not able to efficiently differentiate between the SA and NSA classes as all the values are distributed in the middle of the histogram. On the other hand, the models that included SP efficiently distinguished the SA from the NSA class, because the SA class is concentrated on the right of the histogram, representing a larger probability of self-assembly, while the NSA class is concentrated on the left of the histogram, representing a smaller probability of self-assembly.

PR and ROC curves are plotted for the models that are trained after the hyperparameters are chosen. Examples of PR and ROC curves are included in Figure 4.b. The PR and ROC curves can help determine the ideal threshold value for the binary classification of peptides based on their predicted self-assembly probability. The areas under the curve and the best threshold values are calculated for each of these curves. F1 scores are calculated for points on PR curves and geometric means are calculated for points on ROC curves. These scores can give us additional insight into the performance of the models. In binary classification, the PR curve focuses on the minority class, while the ROC curve covers both classes. ROC curves and ROC AUC can be overly optimistic if the classes are unbalanced. The ROC and PR thresholds that are depicted in Figure 4.a were determined during the hyperparameter optimization process on the peptides used for validation in the inner loop. Thresholds were defined to convert the model output, which is a SA probability between 0 and 1, into a binary class (0 or 1, NSA or SA). The best ROC threshold (shortest distance to the ideal classification threshold of 100% TPR and 0% FPR) and PR threshold (shortest distance to the ideal classification threshold of 100% precision and recall) are indicated by red dots in Figure 4.b. The average values in Figure 4.c were calculated after training all neural network models and predicting the probability of self-assembly for peptides with experimentally validated self-assembly status. The average values are based on all tested random seeds and all tests during hyperparameter optimization.

The neural network model that used only SP values has the highest ROC AUC (0.861), geometric mean (0.788), F1 score (0.848), and accuracy (80.1%) based on the ROC threshold for self-assembly (0.617). It is closely followed

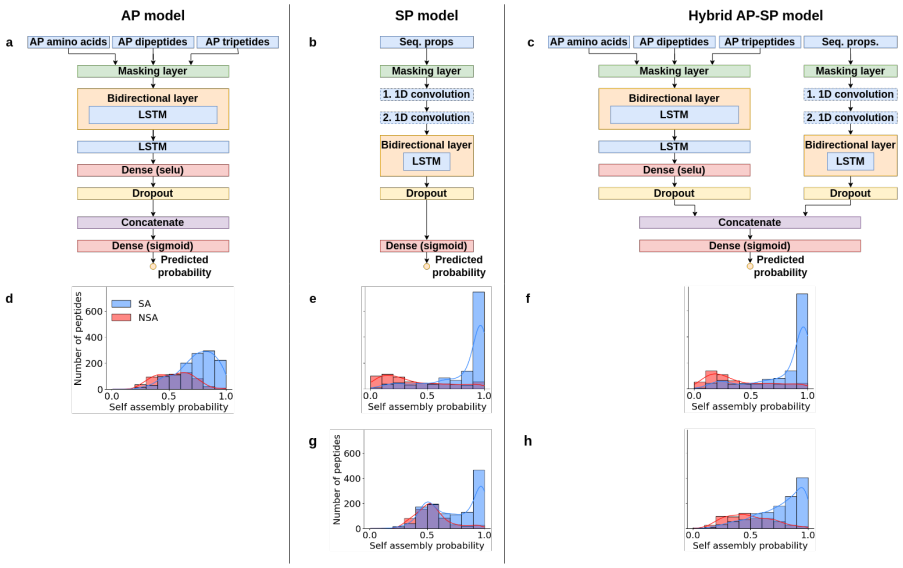


Fig. 3 Architectures and respective performances for AP, SP, and the hybrid AP-SP RNN models. Schematic representation of the RNN models architecture with input data, list of layers, and output (predicted probability) for **a** the AP model; **b** the SP model; and **c** the hybrid AP-SP model. Histograms of predicted self-assembly probability for tested peptide sequences across all tests of **d** the AP model, **e** the SP model, **f** the hybrid AP-SP model, **g** the SP model with t-SNE applied during pre-processing, and **h** the hybrid AP-SP model with t-SNE applied during pre-processing, for experimentally confirmed SA (blue) and NSA (red).

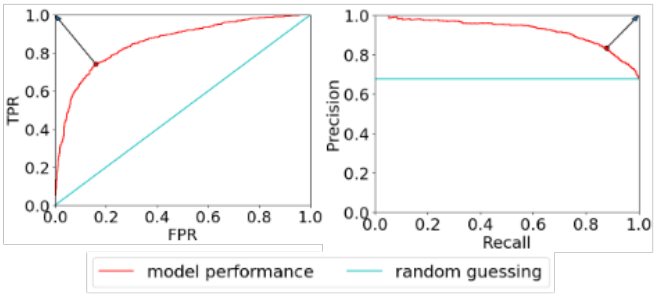
by the hybrid AP-SP model with a ROC AUC of 0.857, a geometric mean of 0.776, an F1 score of 0.841, and an accuracy of 79.1% (the ROC threshold for the AP-SP model is 0.607). Furthermore, the SP neural network model has the largest geometric mean (0.704), F1 score (0.860), and accuracy (79.7%) based on the PR threshold for self-assembly (0.321). The performance of the SP model is closely followed by the hybrid AP-SP model with a geometric mean of 0.692, an F1 score of 0.850, and an accuracy of 78.3% (the PR threshold for the AP-SP model is 0.320). The hybrid AP-SP model has a slightly larger AUC (0.924) than the SP model (0.917) based on the PR curve. Furthermore, the neural network model that used only SP values has the largest geometric mean (0.766), F1 score (0.856), and accuracy (80.4%) based on a threshold for self-assembly of 0.5. It is finally closely followed by the hybrid AP-SP model with a geometric mean of 0.758, an F1 score of 0.849, and an accuracy of 79.5% based on a threshold for self-assembly of 0.5. The 0.5 threshold was used as a standard value without tuning since it splits the possible range of self-assembly probabilities between 0 and 1 exactly in half.

These conclusions match those achieved by measuring the average accuracy and loss for all neural network models across all training epochs based on all tested random seeds and all tests during hyperparameter optimization, as seen

a

Thr	AP	SP	AP-SP	t-SNE SP	t-SNE AP-SP
ROC thr	0.526	0.617	0.607	0.557	0.549
PR thr	0.224	0.321	0.320	0.277	0.309

b



c

Metric	AP	SP	AP-SP	t-SNE SP	t-SNE AP-SP
ROC AUC	0.794	**0.861**	0.857	0.760	0.827
gmean (ROC thr)	0.620	**0.788**	0.776	0.670	0.727
F1 (ROC thr)	0.811	**0.848**	0.841	0.754	0.829
Acc (ROC thr)	72.8%	**80.1%**	79.1%	68.5%	76.8%
PR AUC	0.888	**0.917**	**0.924**	0.885	0.907
gmean (PR thr)	0.044	**0.704**	0.692	0.062	0.463
F1 (PR thr)	0.808	**0.860**	0.850	0.805	0.827
Acc (PR thr)	67.8%	**79.7%**	78.3%	67.4%	72.6%
gmean (0.5)	0.598	**0.766**	0.758	0.582	0.703
F1 (0.5)	0.816	**0.856**	0.849	0.779	0.838
Acc (0.5)	73.0%	**80.4%**	79.5%	68.6%	77.2%

Fig. 4 SA threshold and performance estimation for a set of test peptides. a The average best receiver operating characteristic (ROC) and precision-recall (PR) thresholds for self-assembly for all RNN models. **b** Example ROC curve (left) and PR curve (right) for the final instance of the hybrid AP-SP model trained on all peptide sequences. **c** Performance metrics for the examined RNN models encompassing the average AUC, the geometric mean of TPR and TNR, F1 score, and classification accuracy for the ROC, PR, and standard 0.5 thresholds.

in Figure 2.f, where the SP model has the highest accuracy and the lowest loss during training.

Since the geometric mean exhibits the highest values for all models when using the ROC SA threshold compared to the PR SA threshold and a SA threshold of 0.5, it is considered the best option for the remaining analyses. Although this was not the case with accuracy, which has the highest values with the PR threshold, the geometric mean metric provides a better performance estimate, especially in the case of an imbalanced class distribution such as this dataset.

2.3 What to expect from the model: Comparison to the state-of-the-art

Inspired by a recent study that demonstrated the benefits of using machine intelligence for the discovery of self-assembling peptides [10], we tested the performance of our models in a completely new setting derived from their data. First, we used the 20 designed hexapeptides (Figure 5.a) for which MD-based AP scores calculations were provided along with the experimentally confirmed aggregation status (1 or 0). To separate the true positive (TP) and true negative (TN) classes with a minimal number of false positive (FP) and false negative (FN), the AP cut-off value of 1.765 (Figure 5.a) was determined, which also corresponds to the average AP of the hexapeptides for which extensive MD studies were performed. The cut-off value allowed us to estimate the performance of our pre-trained models in a binary classification task for a new dataset of 6578 hexapeptides without experimentally tested aggregation status.

Figure 5.b presents the same metrics as Figure 4.c for a simpler comparison with previous results along the correlation coefficient analyses between the computed AP value and the predicted SA probability. The relationship between predicted self-assembly probability and AP is further visualized by the scatter plots with a trend line representing linear regression in Figure 5.c, which provide a deeper insight into the impact of the heterogeneous data used to train these models. Spearman and Pearson coefficients of correlation are used to evaluate whether the predicted self-assembly probability is correlated with the AP scores of the peptides, as shown in Figure 5.b. Both Pearson and Spearman coefficients may take a value within the range [-1, 1], with values away from 0 indicating a stronger relationship. The scatter plots, in addition to the Spearman and Pearson coefficients of correlation, show the predicted probability of self-assembly obtained from the AP model is more closely correlated to the AP scores obtained from MD than the predicted probability of self-assembly obtained from any of the other models. This is to be expected because the other models included additional input besides AP in the form of SP values. It is important to note that all models were trained on the small dataset of 158 peptides with known SA propensity that included the 20 hexapeptides for which extensive MD calculations were performed, but none of the remaining 6578 hexapeptides were used to train the model. This provides not only a challenging setting but also a real-world scenario of model exploitation.

Overall, the AP model is the dominant predictor when comparing the AUC ($ROC\ AUC = 0.835$, $PR\ AUC = 0.862$) and correlation scores ($Pearson = 0.60$ and $Spearman = 0.66$). This outcome is expected since this model is trained on AP values of amino acids, dipeptides, and tripeptides and the prediction output is related to AP values of slightly longer sequences of hexapeptides. The correlation level may be interpreted as moderate, revealing that a better prediction requires more than just the AP values of the subsets of the constituting sequence. When comparing the choice of thresholds for

a

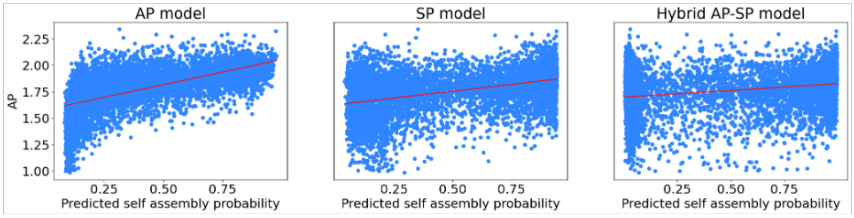
	Peptide	Expert	Agg	AP
1	PTPCY	AI	0	2.54
2	PPPHY	AI	0	2.46
3	SYCGY	AI	1	2.42
4	KWMDF	AI	1	2.28
5	FFEKF	AI	1	2.24
6	KWEFY	AI	1	2.2
7	FKFEF	Human	1	2.12
8	RWLDY	AI	1	2.11
9	WKPYY	AI	0	2.11
10	VVVVV	Human	1	2.06
11	FKIDF	AI	1	2.03
12	VKVFF	Human	1	1.95
13	KFFFE	Human	1	1.95
14	KFAFD	Human	1	1.86
15	VKVEV	Human	1	1.84
Cut-off = 1.765				
16	RVSVD	Human	0	1.66
17	KKFDD	Human	0	1.64
18	VKVKV	Human	0	1.35
19	KVKVK	Human	0	1.19
20	DPDPD	Human	0	1.01

TP FP
FN TN

b

Metric	AP	SP	AP-SP	t-SNE SP	t-SNE AP-SP
ROC AUC	**0.835**	0.716	0.641	0.627	0.808
gmean (ROC thr)	0.650	0.669	0.614	0.548	**0.701**
F1 (ROC thr)	0.604	**0.684**	0.640	0.502	**0.683**
Acc (ROC thr)	65.7%	66.5%	61.3%	55.2%	**69.5%**
PR AUC	**0.862**	0.753	0.719	0.674	0.834
gmean (PR thr)	**0.743**	0.650	0.592	0.000	0.714
F1 (PR thr)	**0.803**	0.729	0.669	0.726	0.781
Acc (PR thr)	**76.3%**	67.5%	61.1%	57.0%	73.6%
gmean (0.5)	0.666	0.674	0.610	0.578	**0.715**
F1 (0.5)	0.627	**0.708**	0.654	0.579	**0.710**
Acc (0.5)	66.9%	67.6%	61.4%	57.2%	**70.8%**
Pearson	**0.60**	0.42	0.26	0.23	0.56
Spearman	**0.66**	0.42	0.27	0.23	0.59

c



d

Model	Attempts	gmean	F1	Acc
Human [10]	11	0.000	0.706	55%
AI [10]	9	0.000	0.800	67%
AP	11+9	**0.685**	0.750	70%
SP	11+9	0.677	**0.815**	**75%**
AP-SP	11+9	0.645	0.769	70%
t-SNE SP	11+9	0.000	0.750	60%
t-SNE AP-SP	11+9	0.645	0.769	70%

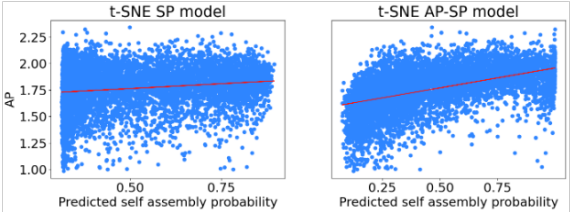


Fig. 5 Performance estimation for a set of 6578 hexapeptides analyzed by MD. **a** The 20 hexapeptides with their experimentally confirmed aggregation status are sorted in descending order by their AP calculated with MD. Having estimated a cut-off value for binary classification, the instances are color coded using the confusion matrix outcomes: TP (green), TN (yellow), FP (red), FN (blue). **b** Table of performance metrics for all neural network models, containing the AUC, *gmean*, F1 score, and accuracy for every threshold (ROC, PR, and standard 0.5) in a separated sub-table. The last sub-table contains the Pearson and Spearman correlation coefficients for the AP - SA relationship. **c** Scatter plots depicting the AP-SA relationship. The red line represents a linear regression function for the given values. **d** Geometric mean, F1 score, and accuracy using the PR threshold for all created models and the reference for the human and AI peptides.

the final classification, the AP model is consistent with the previous analysis and produces the best performance using the PR threshold ($gmean = 0.743$, $F1 = 0.803$, and $Acc = 76.3\%$). In contrast, the t-SNE AP-SP model performs best with the ROC threshold ($gmean = 0.701$, $F1 = 0.683$ and $Acc = 69.5\%$), along with the SP model ($F1 = 0.684$). In addition to proving that the choice of optimal threshold is crucial for obtaining the best-performing classifier, this also shows that the SP values contain information of utmost importance in revealing the complex sequence-structure relationship. Reducing the number of

SP values using the t-SNE technique allowed the model to obtain the second-best correlation levels ($Pearson = 0.56$, $Spearman = 0.59$) and AUC values ($ROC\ AUC = 0.808$, $PR\ AUC = 0.834$) after the AP model, and the best performance with the standard threshold ($gmean = 0.715$, $F1 = 0.710$ and $Acc = 70.8\%$), closely followed by the SP model ($F1 = 0.708$).

Having a hybrid t-SNE AP-SP model proved beneficial in predicting the SA propensity in some analyses, while it is outperformed by the simpler AP model and the SP model in other scenarios. Dimensionality reduction resulted in an equal number of features that came from physico-chemical properties and MD simulations, allowing for a simpler model that can generalize knowledge in a more stable manner. Although the evaluation results are not the same as those of peptides with experimentally validated self-assembly, the model that combines heterogeneous data has the advantage of combining AP and SP data as well as peptide sequences from multiple sources that enable more extensive testing.

Experts predicted the ability to self-assemble for 11 peptides, and artificial intelligence predicted it for an additional 9 sequences using the Batra et al. model [10], bringing the total number of hexapeptide sequences extensively studied by MD to 20. Our five models predicted the probability of self-assembly for both sets of sequences, and the PR threshold for each model (as defined in Figure 4.a) was applied to convert the predictions to binary classes. Metrics calculated based on these predictions are visible in Figure 5.d. Human-based self-assembly prediction for 11 sequences had an accuracy of 55%. Our five models were more successful in terms of accuracy on the 20 sequences used. On the other hand, all of our five models showed a higher accuracy on 20 sequences, except for the t-SNE SP model, which has an accuracy of 60%, compared to the Batra et al. model, which had an accuracy of 67% on 9 hexapeptides. Since the human predictions and the predictions of the Batra et al. model predicted only the positive class, the true-negative rate (TNR) and the false-negative rate (FNR) are equal to 0, the geometric mean is also equal to 0, while the accuracy is equal to the TPR. From these values, the geometric mean and F1 results were estimated and compared with our five models. All of our five models outperform human predictions based on their F1 score. The SP model outperforms the predictions of the Batra et al. model based on F1 scores. All of our five models have the same or better geometric mean results than human predictions and the predictions of the Batra et al. model.

3 Methods

All code was written in the *Python* programming language.

The sliding window approach. In models that use AP values (amino acids, di-, and tri- peptides as subsets of the peptide sequence), they are obtained using a sliding window of the appropriate size and stored in an array to be used as input for the model. When amino acids are used, the array containing the calculated values is of the same length as the peptide. For di- or

tri- peptides, the arrays are shorter than the peptide sequence by one or two entries. The same approach is used for each SP value for models based on SP [47], but in this case, the sliding window size is always 1, and only individual amino acids are considered.

One of the architectures used only the AP values of amino acids, dipeptides, or tripeptides in three submodels. Another model architecture that was tested uses SP values as input for a single model. AP and SP values are used as submodels for hybrid AP-SP model architectures. Other features relevant to AP, such as $logP$ and AP_H [3] were tested for the hybrid model, but did not improve its performance.

Scaling input values. AP values for amino acids [46], dipeptides [57] and tripeptides [3] were obtained from published articles. AP scores of di- and tripeptides represent the percentage of peptide surface exposed to water before and after aggregation and have values of 1 if no structure is formed or values greater than 1 in case of aggregation. On the other hand, amino acid AP values represent the energy released during the formation of supramolecular structures, can assume negative values, making them not directly comparable to AP scores of dipeptides and tripeptides. Normalization of input data was performed to bring diverse input features to a comparable value range, facilitating gradient flow, yielding a model with superior performance [58, 59]. In our study, we used Min-Max scaling to map the input data to a range $[-1, 1]$.

As previously reported, 94 physico-chemical properties were used for each amino acid in the peptide sequence and stored in a matrix for each peptide sequence [47]. The matrix has a number of rows equal to the length of the peptide, and a number of columns equal to the number of properties. When peptides are processed in batches, all of the peptides are padded to have a length equal to the longest sequence in the batch. All values are scaled to fit into a range that can be modified, but the default range of values used was between -1 and 1 because it is preferable to center the range around 0 for machine learning models. The default padding value used was 0, but this was replaced with 2 so that the padding value would be outside the range used for the relevant data. Once the padded values are outside the range used for relevant data, they can be masked for processing in ML models to ensure that padding is ignored while training neural networks and does not impact model weights.

Initial settings and callbacks during training. Training was limited to 70 epochs. The batch size was fixed at 600 to ensure that all peptides are processed in a single batch to obtain the fastest speed of operation and smoother gradients. The learning rate starts at 0,01. A custom function keeps the initial learning rate for the first ten epochs and decreases it exponentially afterward by multiplying the learning rate by $e^{0.1}$. Only the model with the lowest validation loss is saved.

Hyperparameter optimization using nested k-fold cross-validation. Nested k-fold cross-validation splits the original dataset into an outer (i) training and validation fold and an outer (ii) testing fold. The outer (i) training and validation fold is later split again into an (iii) inner training and validation fold and (iv) inner testing fold for hyperparameter tuning. A model is trained with each inner training fold with all possible combinations of hyperparameters chosen for the grid search. The hyperparameters that yield a model with the lowest average validation loss over all of the inner validation folds are applied in training and testing the model with the outer folds. The parameter k was set to 5, indicating a thorough optimization process that yields 5 repeated measurements, as presented in figure 2.d.

Input dimension and layers for different models. The input dimensions for the submodel that includes SP as input are (None, 24, 94) because there can be any number of sequences, all padded to a length of 24 with 94 SP values encoded for each amino acid in the sequence. The input dimensions for the three submodels that include only AP values as input are (None, 24, 1) because there can be any number of sequences all padded to a length of 24 with one array representing AP values for amino acids, dipeptides, or tripeptides which are subsets of the peptide sequence.

All values are scaled to fit into a range between -1 and 1. Padding is used to ensure that all input vectors to the model are the same size so that a batch size greater than 1 can be used for model training, validation, and prediction. The default padding value used is 2 because it does not belong to the range between -1 and 1.

All models apply a masking layer so that the padded values (in this case 2) that were added to ensure that all input vectors have the same length are ignored.

Dropout, concatenating submodels, and final prediction. A dropout value of 0.5 is applied in the final layer for all the submodels used. A large dropout is needed to prevent overfitting because a complex model is applied to a small data set. Dropout values of 0.1, 0.2, and 0.3 were also tested but were not part of hyperparameter optimization. In models that use multiple submodels, a concatenation layer is added that merges multiple arrays into one before the final prediction is made. A final layer with sigmoid activation is applied to ensure that the model generates a single value between 0 and 1 in the output of each peptide in the input data representing the probability of self-assembly.

Dataset. Sequences longer than 50 amino acids were excluded from this analysis as they are considered proteins. Peptide sequences that have been tested to determine whether they exhibit self-assembly were obtained from different sources, including [60], which contained 91 peptides without self-assembly and 67 peptides with self-assembly.

4 Conclusion

Longer peptides are made up of simpler building blocks (amino acids, dipeptides, and tripeptides) whose properties might be useful in determining the properties of the peptide sequence as a whole. The exponentially increasing number of combinations of 20 amino acids poses a challenge when examining these longer peptides. The appearance of RNNs, which work well with sequential data, enabled us to create a prediction model of the self-assembly propensity of peptides employing only their sequences as input.

AP values calculated by MD in extensive experiments were used along with SP values. AP is an important factor in determining the behavior of longer peptide structures, but it is known that other properties influence self-assembly as well. This was the motivation for introducing additional variables in the form of SP values as input for some of the developed RNN models in order to achieve an efficient prediction of AP for longer sequences.

In order to utilize the available information about the AP of amino acids, dipeptides, and tripeptides as well as SP values, different RNN models were developed using heterogeneous data to build a comprehensive knowledge-based model. One of the model configurations used only AP values, another used only SP values, and a hybrid AP-SP model used both. Two additional models were created after reducing the number of SP values using t-SNE, one of which used only SP values whose number was reduced using t-SNE, and another which used AP values and SP values whose number was reduced using t-SNE.

Good results were obtained both for peptides that exhibit and peptides that do not exhibit self-assembly, as proven by high geometric mean scores (0.704 for the SP model and PR threshold) that are very close to accuracy scores (79.7% for the SP model and PR threshold). The SP model, when used with the PR threshold, outperforms human predictions and the predictions of the state-of-the-art model from the literature based on accuracy, F1 score, and geometric mean.

Declarations

- Code availability
  The code is available on a public GitHub repository: https://github.com/LucijaZuzic/peptide_properties.git.

Appendix A Structure of the input data for an example sequence for different neural network models

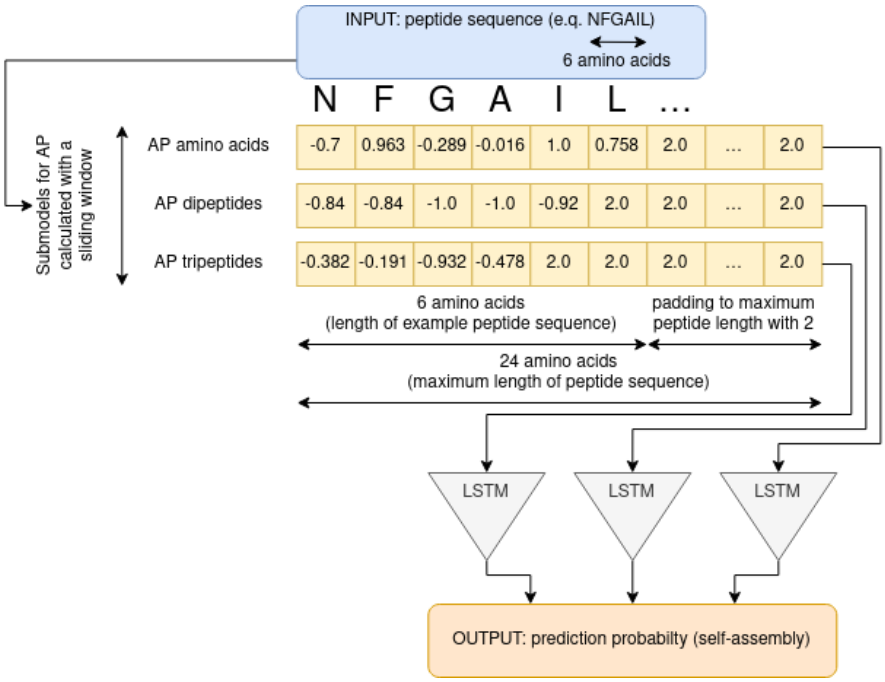


Fig. A1 Structure of the input data for the example sequence (NFGAIL) for an AP neural network model.

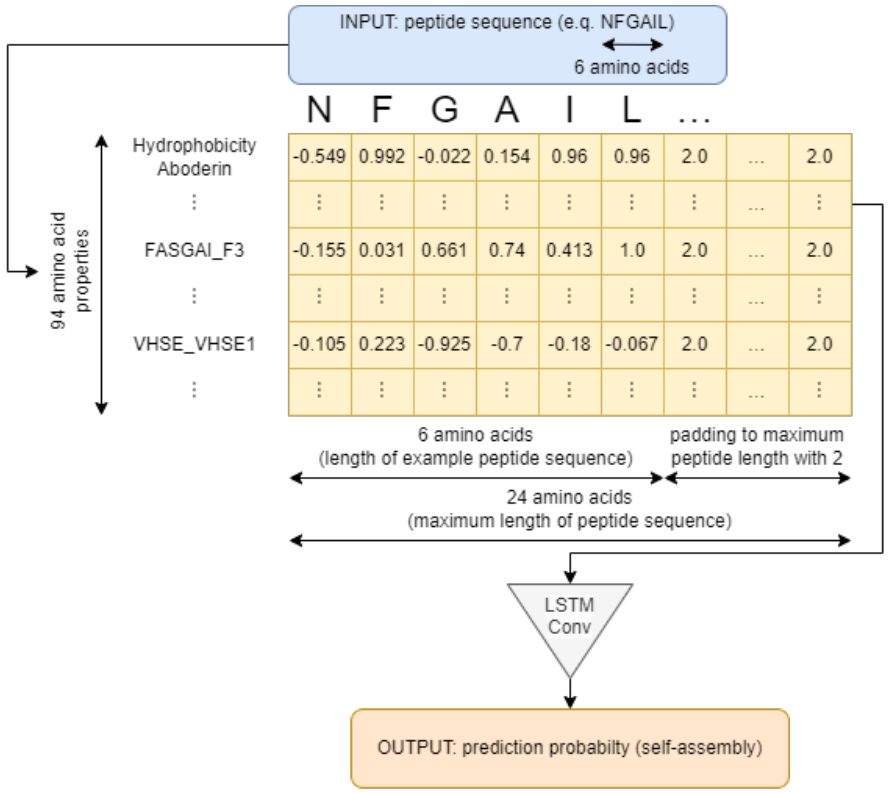


Fig. A2 Structure of the input data for the example sequence (NFGAIL) for an SP neural network model.

Appendix B Example ROC curves and PR curves for different neural network models

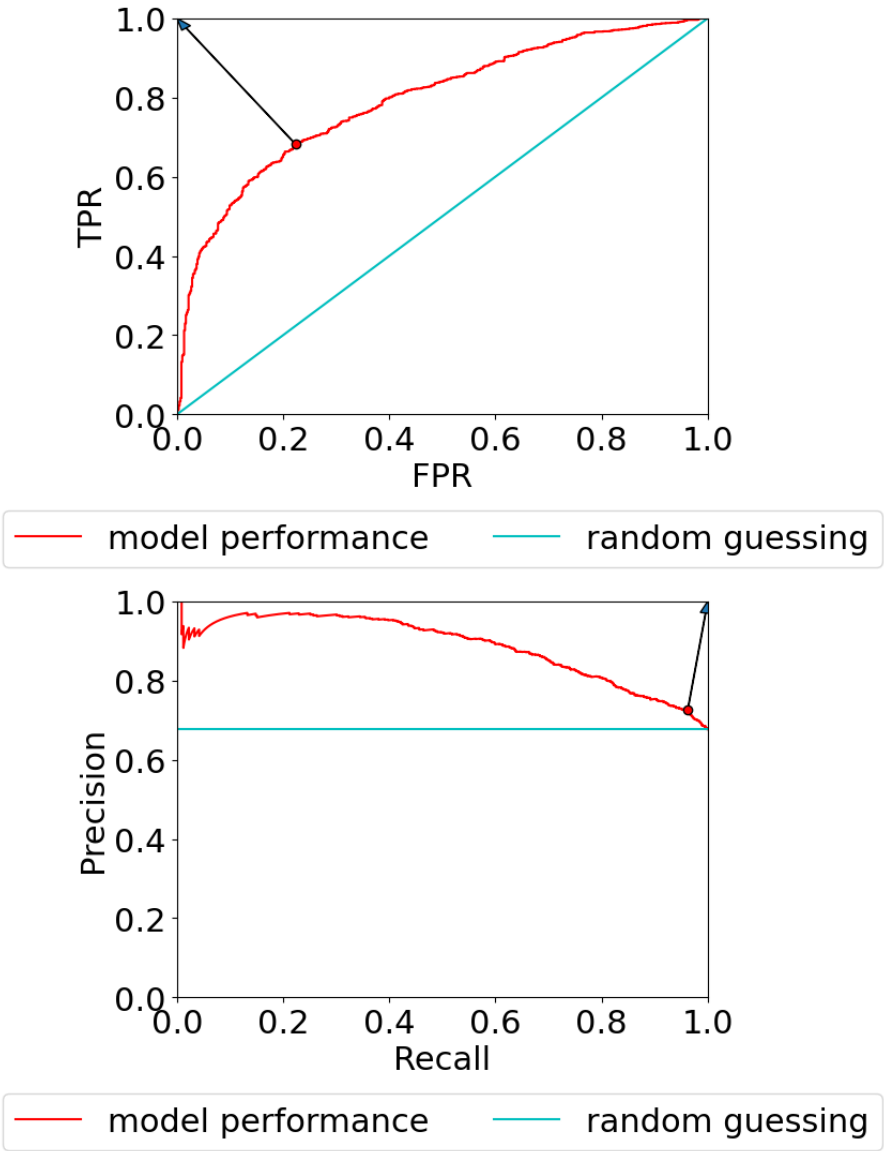


Fig. B3 Example ROC curve (top) and PR curve (bottom) for the final instance of the AP model trained on all peptide sequences.

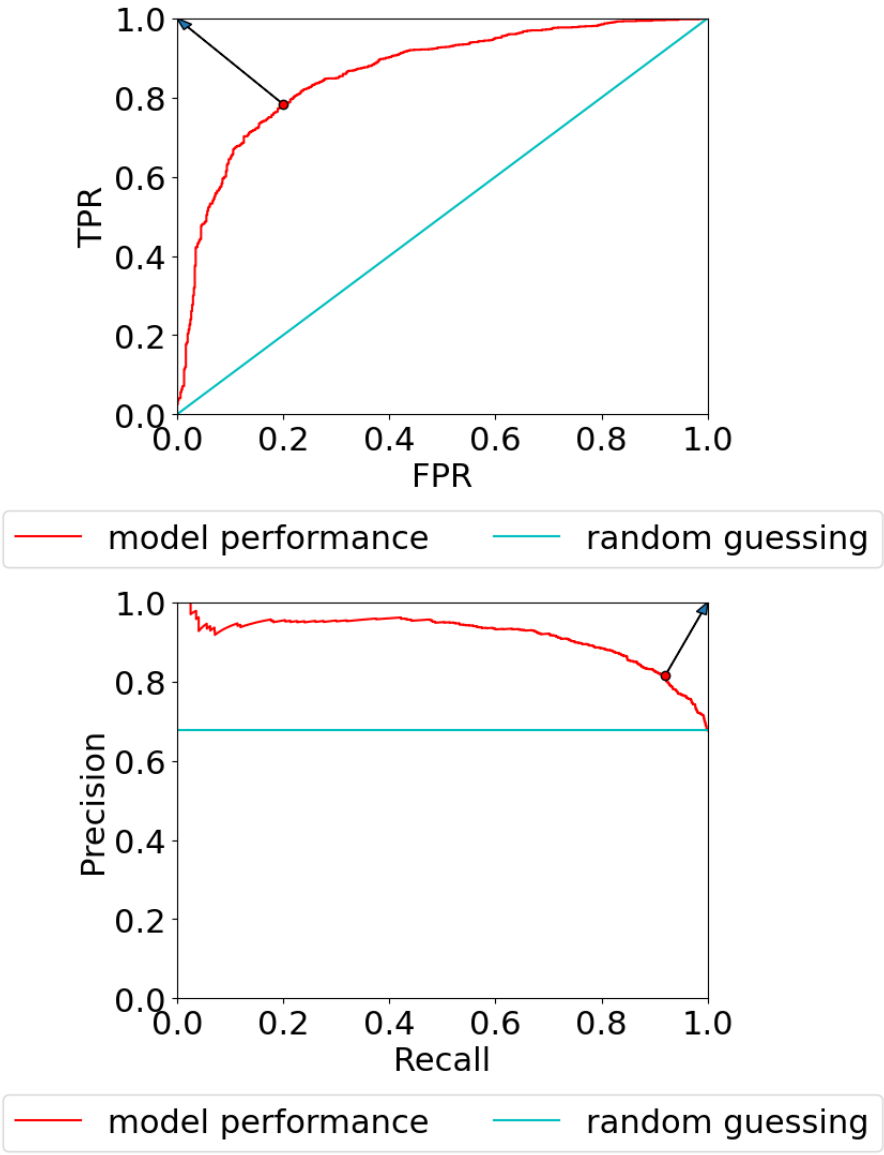


Fig. B4 Example ROC curve (top) and PR curve (bottom) for the final instance of the SP model trained on all peptide sequences.

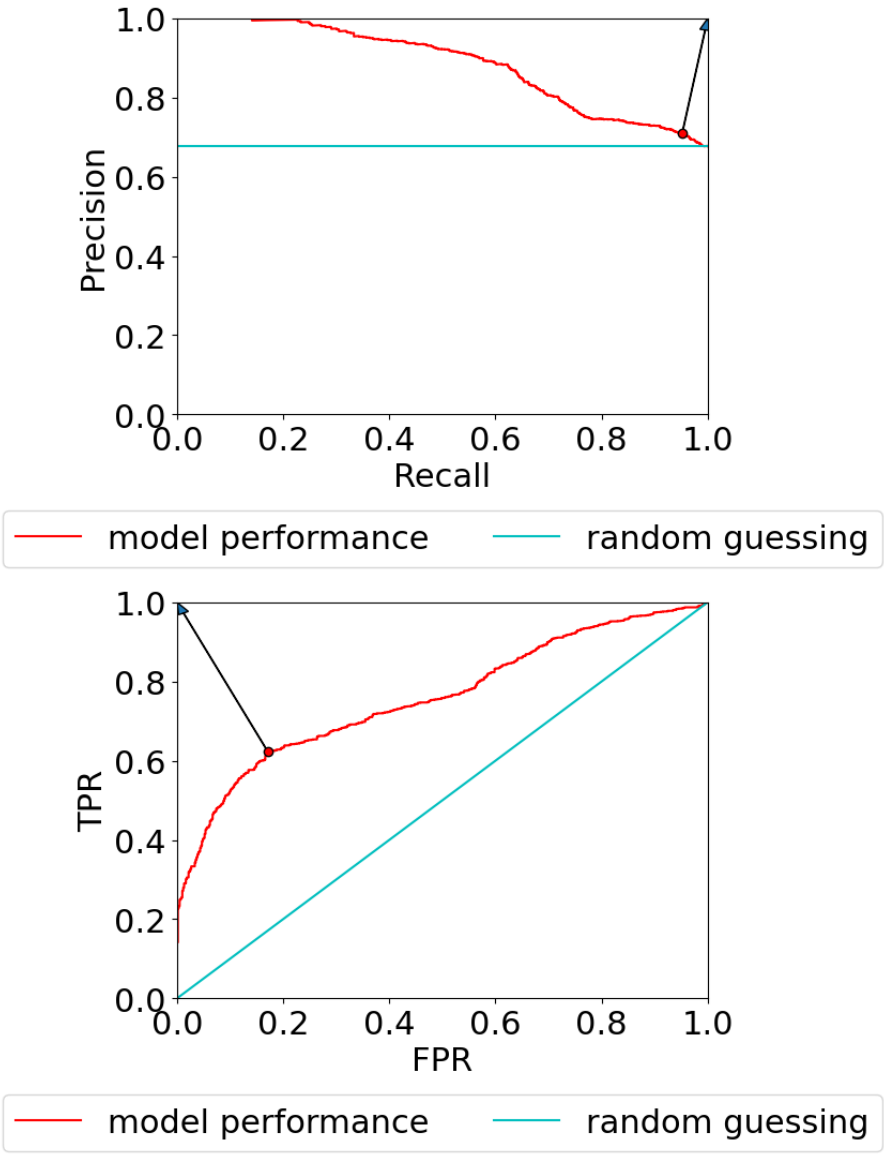


Fig. B5 Example ROC curve (top) and PR curve (bottom) for the final instance of the t-SNE SP model trained on all peptide sequences.

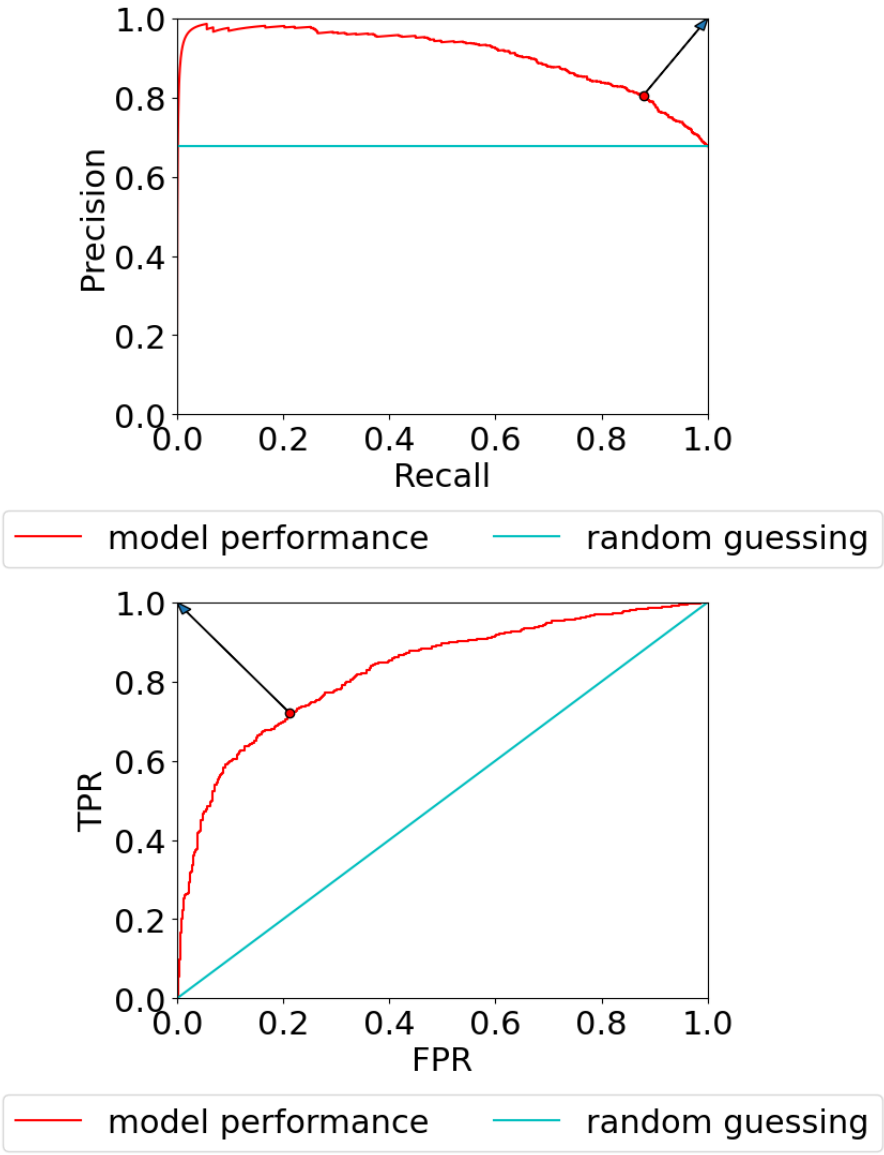


Fig. B6 Example ROC curve (top) and PR curve (bottom) for the final instance of the t-SNE AP-SP model trained on all peptide sequences.

References

[1] Lampel, A.: Biology-inspired supramolecular peptide systems. Chem **6**(6), 1222–1236 (2020)

[2] Janković, P., Šantek, I., Pina, A.S., Kalafatovic, D.: Exploiting peptide self-assembly for the development of minimalistic viral mimetics. Frontiers in Chemistry **9**, 723473 (2021)

[3] Frederix, P.W., Scott, G.G., Abul-Haija, Y.M., Kalafatovic, D., Pappas, C.G., Javid, N., Hunt, N.T., Ulijn, R.V., Tuttle, T.: Exploring the sequence space for (tri-) peptide self-assembly to design and discover new hydrogels. Nature chemistry **7**(1), 30–37 (2015)

[4] Lampel, A., Ulijn, R., Tuttle, T.: Guiding principles for peptide nanotechnology through directed discovery. Chemical Society Reviews **47**(10), 3737–3758 (2018)

[5] Levin, A., Hakala, T.A., Schnaider, L., Bernardes, G.J., Gazit, E., Knowles, T.P.: Biomimetic peptide self-assembly for functional materials. Nature Reviews Chemistry **4**(11), 615–634 (2020)

[6] Chatterjee, A., Reja, A., Pal, S., Das, D.: Systems chemistry of peptide-assemblies for biochemical transformations. Chemical Society Reviews (2022)

[7] Ramakrishnan, M., van Teijlingen, A., Tuttle, T., Ulijn, R.V.: Integrating computation, experiment, and machine learning in the design of peptide-based supramolecular materials and systems. Angewandte Chemie (2023)

[8] Lampel, A., McPhee, S.A., Park, H.-A., Scott, G.G., Humagain, S., Hekstra, D.R., Yoo, B., Frederix, P.W., Li, T.-D., Abzalimov, R.R., *et al.*: Polymeric peptide pigments with sequence-encoded properties. Science **356**(6342), 1064–1068 (2017)

[9] Smith, D.J., Brat, G.A., Medina, S.H., Tong, D., Huang, Y., Grahammer, J., Furtmüller, G.J., Oh, B.C., Nagy-Smith, K.J., Walczak, P., *et al.*: A multiphase transitioning peptide hydrogel for suturing ultrasmall vessels. Nature nanotechnology **11**(1), 95–102 (2016)

[10] Batra, R., Loeffler, T.D., Chan, H., Srinivasan, S., Cui, H., Korendovych, I.V., Nanda, V., Palmer, L.C., Solomon, L.A., Fry, H.C., et al.: Machine learning overcomes human bias in the discovery of self-assembling peptides. Nature chemistry, 1–9 (2022)

[11] Pierce, N.A., Winfree, E.: Protein design is np-hard. Protein engineering **15**(10), 779–782 (2002)

[12] Hu, K., Xiong, W., Sun, C., Wang, C., Li, J., Yin, F., Jiang, Y., Zhang, M.-R., Li, Z., Wang, X., *et al.*: Self-assembly of constrained cyclic peptides controlled by ring size. CCS Chemistry **2**(1), 42–51 (2020)

[13] Hu, K., Jiang, Y., Xiong, W., Li, H., Zhang, P.-Y., Yin, F., Zhang, Q., Geng, H., Jiang, F., Li, Z., *et al.*: Tuning peptide self-assembly by an in-tether chiral center. Science advances **4**(5), 5907 (2018)

[14] Chan, K.H., Lee, W.H., Ni, M., Loo, Y., Hauser, C.A.: C-terminal residue of ultrashort peptides impacts on molecular self-assembly, hydrogelation, and interaction with small-molecule drugs. Scientific reports **8**(1), 1–14 (2018)

[15] Kim, J., Han, T.H., Kim, Y.-I., Park, J.S., Choi, J., Churchill, D.G., Kim, S.O., Ihee, H.: Role of water in directing diphenylalanine assembly into nanotubes and nanowires. Advanced Materials **22**(5), 583–587 (2010)

[16] Nguyen, P.K., Gao, W., Patel, S.D., Siddiqui, Z., Weiner, S., Shimizu, E., Sarkar, B., Kumar, V.A.: Self-assembly of a dentinogenic peptide hydrogel. ACS omega **3**(6), 5980–5987 (2018)

[17] Yan, X., Cui, Y., He, Q., Wang, K., Li, J., Mu, W., Wang, B., Ou-yang, Z.-c.: Reversible transitions between peptide nanotubes and vesicle-like structures including theoretical modeling studies. Chemistry–A European Journal **14**(19), 5974–5980 (2008)

[18] Yang, K.K., Wu, Z., Arnold, F.H.: Machine-learning-guided directed evolution for protein engineering. Nature methods **16**(8), 687–694 (2019)

[19] Mandal, D., Shirazi, A.N., Parang, K.: Self-assembly of peptides to nanostructures. Organic & biomolecular chemistry **12**(22), 3544–3561 (2014)

[20] Shmilovich, K., Mansbach, R.A., Sidky, H., Dunne, O.E., Panda, S.S., Tovar, J.D., Ferguson, A.L.: Discovery of self-assembling π-conjugated peptides by active learning-directed coarse-grained molecular simulation. The Journal of Physical Chemistry B **124**(19), 3873–3891 (2020)

[21] Gocheva, G., Peneva, K., Ivanova, A.: Self-assembly of doxorubicin and a drug-binding peptide studied by molecular dynamics. Chemical Physics **525**, 110380 (2019)

[22] Guo, C., Luo, Y., Zhou, R., Wei, G.: Triphenylalanine peptides self-assemble into nanospheres and nanorods that are different from the nanovesicles and nanotubes formed by diphenylalanine peptides. Nanoscale **6**(5), 2800–2811 (2014)

[23] Lee, O.-S., Cho, V., Schatz, G.C.: Modeling the self-assembly of peptide amphiphiles into fibers using coarse-grained molecular dynamics. Nano letters **12**(9), 4907–4913 (2012)

[24] Hauser, C.A., Deng, R., Mishra, A., Loo, Y., Khoe, U., Zhuang, F., Cheong, D.W., Accardo, A., Sullivan, M.B., Riekel, C., *et al.*: Natural tri-to hexapeptides self-assemble in water to amyloid β-type fiber aggregates by unexpected α-helical intermediate structures. Proceedings of the National Academy of Sciences **108**(4), 1361–1366 (2011)

[25] Frederix, P.W., Patmanidis, I., Marrink, S.J.: Molecular simulations of self-assembling bio-inspired supramolecular systems and their connection to experiments. Chemical Society Reviews **47**(10), 3470–3489 (2018)

[26] Takahashi, K., Oda, T., Naruse, K.: Coarse-grained molecular dynamics simulations of biomolecules. AIMS Biophysics **1**(1), 1–15 (2014)

[27] Frederix, P.W., Ulijn, R.V., Hunt, N.T., Tuttle, T.: Virtual screening for dipeptide aggregation: Toward predictive tools for peptide self-assembly. The journal of physical chemistry letters **2**(19), 2380–2384 (2011)

[28] Zhou, P., Yuan, C., Yan, X.: Computational approaches for understanding and predicting the self-assembled peptide hydrogels. Current Opinion in Colloid & Interface Science, 101645 (2022)

[29] Palmer, N., Maasch, J.R., Torres, M.D., de la Fuente-Nunez, C.: Molecular dynamics for antimicrobial peptide discovery. Infection and Immunity **89**(4), 00703–20 (2021)

[30] Zeng, W.-F., Zhou, X.-X., Willems, S., Ammar, C., Wahle, M., Bludau, I., Voytik, E., Strauss, M.T., Mann, M.: Alphapeptdeep: a modular deep learning framework to predict peptide properties for proteomics. Nature Communications **13**(1), 1–14 (2022)

[31] Bukhari, S.N.H., Webber, J., Mehbodniya, A.: Decision tree based ensemble machine learning model for the prediction of zika virus t-cell epitopes as potential vaccine candidates. Scientific Reports **12**(1), 1–11 (2022)

[32] Melo, M.C., Maasch, J.R., de la Fuente-Nunez, C.: Accelerating antibiotic discovery through artificial intelligence. Communications biology **4**(1), 1050 (2021)

[33] Chen, J., Cheong, H.H., Siu, S.W.: Xdeep-acpep: deep learning method for anticancer peptide activity prediction based on convolutional neural network and multitask learning. Journal of chemical information and modeling **61**(8), 3789–3803 (2021)

[34] Akbar, S., Ahmad, A., Hayat, M., Rehman, A.U., Khan, S., Ali, F.: iatbp-hyb-enc: Prediction of antitubercular peptides via heterogeneous feature representation and genetic algorithm based ensemble learning model. Computers in Biology and Medicine **137**, 104778 (2021)

[35] Aronica, P.G., Reid, L.M., Desai, N., Li, J., Fox, S.J., Yadahalli, S., Essex, J.W., Verma, C.S.: Computational methods and tools in antimicrobial peptide research. Journal of Chemical Information and Modeling **61**(7), 3172–3196 (2021)

[36] Hasan, M.M., Schaduangrat, N., Basith, S., Lee, G., Shoombuatong, W., Manavalan, B.: Hlppred-fuse: improved and robust prediction of hemolytic peptide and its activity by fusing multiple feature representation. Bioinformatics **36**(11), 3350–3356 (2020)

[37] Manavalan, B., Shin, T.H., Kim, M.O., Lee, G.: Aippred: sequence-based prediction of anti-inflammatory peptides using random forest. Frontiers in pharmacology **9**, 276 (2018)

[38] Attique, M., Farooq, M.S., Khelifi, A., Abid, A.: Prediction of therapeutic peptides using machine learning: computational models, datasets, and feature encodings. IEEE Access **8**, 148570–148594 (2020)

[39] Li, F., Han, J., Cao, T., Lam, W., Fan, B., Tang, W., Chen, S., Fok, K.L., Li, L.: Design of self-assembly dipeptide hydrogels and machine learning via their chemical features. Proceedings of the National Academy of Sciences **116**(23), 11259–11264 (2019)

[40] van Teijlingen, A., Tuttle, T.: Beyond tripeptides two-step active machine learning for very large data sets. Journal of Chemical Theory and Computation **17**(5), 3221–3232 (2021)

[41] Scott, G.G., Börner, T., Leser, M.E., Wooster, T.J., Tuttle, T.: Directed discovery of tetrapeptide emulsifiers. Frontiers in chemistry **10** (2022)

[42] Heydari, S., Raniolo, S., Livi, L., Limongelli, V.: Transferring chemical and energetic knowledge between molecular systems with machine learning. Communications Chemistry **6**(1), 13 (2023)

[43] Bengio, Y., Simard, P., Frasconi, P.: Learning long-term dependencies with gradient descent is difficult. IEEE transactions on neural networks **5**(2), 157–166 (1994)

[44] Yang, G., Jiayu, Y., Dongdong, X., Zelin, G., Hai, H.: Feature-enhanced text-inception model for chinese long text classification. Scientific Reports **13**(1), 2087 (2023)

[45] Zhang, L., Wang, S., Liu, B.: Deep learning for sentiment analysis: A survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery **8**(4), 1253 (2018)

[46] De Groot, N., Pallarès, I., Avilés, F., Vendrell, J., Ventura, S.: Prediction of hot spots of aggregation in disease-linked polypeptides,. BMC structural biology **5**, 18 (2005)

[47] Otović, E., Njirjak, M., Kalafatovic, D., Mauša, G.: Sequential properties representation scheme for recurrent neural network-based prediction of therapeutic peptides. Journal of Chemical Information and Modeling **62**(12), 2961–2972 (2022)

[48] Conchillo-Solé, O., De Groot, N., Avilés, F., Vendrell, J., Daura, X., Ventura, S.: AGGRESCAN: a server for the prediction of "hot spots" of aggregation in polypeptides. BMC bioinformatics **8**, 65 (2007)

[49] Lee, S., Trinh, T.H., Yoo, M., Shin, J., Lee, H., Kim, J., Hwang, E., Lim, Y., Ryou, C.: Self-Assembling Peptides and Their Application in the Treatment of Diseases. International Journal of Molecular Sciences **20** (2019)

[50] team, A.S.: Hey siri: An on-device dnn-powered voice trigger for apple's personal assistant. Machine Learning Research at Apple **0** (2017)

[51] Le, Q.V., Schuster, M.: A neural network for machine translation, at production scale. Google AI Blog **27** (2016)

[52] McNemar, Q.: Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika **12**(2), 153–157 (1947)

[53] Pascanu, R., Mikolov, T., Bengio, Y.: On the difficulty of training recurrent neural networks. In: International Conference on Machine Learning, pp. 1310–1318 (2013). Pmlr

[54] Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(11) (2008)

[55] Wei, L., Ye, X., Sakurai, T., Mu, Z., Wei, L.: Toxibtl: prediction of peptide toxicity based on information bottleneck and transfer learning. Bioinformatics **38**(6), 1514–1524 (2022)

[56] Dean, S.N., Alvarez, J.A.E., Zabetakis, D., Walper, S.A., Malanoski, A.P.: Pepvae: variational autoencoder framework for antimicrobial peptide generation and activity prediction. Frontiers in Microbiology **12**, 725727 (2021)

[57] Frederix, P., Ulijn, R., Hunt, N., Tuttle, T.: Virtual Screening for Dipeptide Aggregation: Toward Predictive Tools for Peptide Self-Assembly,. The journal of physical chemistry letters **2**, 2380–2384 (2011)

[58] Singh, D., Singh, B.: Investigating the impact of data normalization on classification performance. Applied Soft Computing **97**, 105524 (2020)

[59] Nawi, N.M., Atomi, W.H., Rehman, M.Z.: The effect of data preprocessing on optimized training of artificial neural networks. Procedia Technology **11**, 32–39 (2013)

[60] Thompson, M., Sievers, S., Karanicolas, J., Ivanova, M., Baker, D., Eisenberg, D.: The 3D profile method for identifying fibril-forming segments of proteins. Proceedings of the National Academy of Sciences of the United States of America **103**, 4074–8 (2006)