ĐẠI HỌC QUỐC GIA HÀ NỘI

TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

TIỂU LUẬN

Ngành Kĩ thuật điện tử và Tin học

ĐỀ TÀI:

Học máy và bài toán năng lượng liên kết phối tử – thụ thể

Sinh viên: Vũ Trường Thành MSV: 21002235

Giáo viên hướng dẫn : GS.TS. Nguyễn Thế Toàn

HàNội, 2023

MỤC LỤC

LỜI CẢM ƠN

Để có thể hoàn thành tiểu luận về đề tài: Học máy và bài toán năng lượng liên kết phối tử - thụ thể là kết quả của quá trình học tập, tiếp thu kiến thức tại trường, lớp và cả những tìm tòi, nghiên cứu riêng của bản thân em và sự chỉ dạy tận tình của thầy Nguyễn Thế Toàn, các anh chị Keylab và thầy/cô khoa vật lý - người đã trực tiếp hướng dẫn em trong môn học này. Do vậy, qua đây em xin cảm ơn các thầy, cô, các bạn và anh chị.

Mặc dù đã dành thời gian và nỗ lực để hoàn thành bài tiểu luận này, nhưng do sự hạn chế về mặt kiến thức nên bài làm khó tránh khỏi những thiếu sót. Em kính mong nhận được những lời góp ý của quý thầy, cô để bài làm ngày càng hoàn thiện hơn. Em xin chân thành cảm ơn!

Danh sách hình vẽ

Mở đầu

1. Lí do chọn đề tài.

Học máy (Machine learning) được biết đến là một phần quan trọng trong ngành *trí tuệ nhân tạo AI* và *khoa học máy tính* tập trung vào việc thu thập, phân tích, sử dụng dữ liệu và thuật toán để bắt trước cách con người học, dần dần cải thiện độ chính xác của nó. Hàng loạt các ứng dụng sử dụng machine learning ra đời trên mọi linh vực của cuộc sống, từ khoa học máy tính đến những ngành ít liên quan hơn như vật lý, hóa học, y học, chính trị.

Khám phá phối tử dựa trên cấu trúc là một trong những phương pháp thành công nhất để tăng cường quá trình khám phá thuốc. Hiện tại, có một sự thay đổi đáng chú ý đối với các phương pháp học máy để hỗ trợ các quy trình đó. Học sâu gần đây đã thu hút được sự chú ý đáng kể vì nó cho phép mô hình 'học' để trích xuất các tính năng phù hợp với nhiệm vụ hiện tại.

Vì vậy bài tiểu luận sẽ là bước mở đầu để em bắt đầu tìm hiểu và làm quen trong lĩnh vực về học máy với vật lý y sinh. Bắt đầu với bài toán: c*ho trước cấu trúc complex của ligand và protein, làm thế nào để tính ra năng lượng tương tác giữa hai molecules?*

2. Mục tiêu.

Với đề tài này, em mong muốn được bắt đầu tiếp cận làm quen và học hỏi với học máy. Qua đó có kiến thức đẻ tìm hiểu các bài báo nghiên cứu và trau dồi thêm kiến thức để dần xây dựng nên các bài nhiên cứu cá nhân.

3. Phương pháp.

Tra cứu tài liêu,các bài báo quốc tế, tổng hợp và phân tích thông tin, nghiên cứu và thực nghiệm, đưa ra những nhận xét, đánh giá.

Nội dung

Bài toán tính năng lượng tương tác giữa hai *molecules* từ cấu trúc phức hợp của *phối tử và protein* cho trước có thể sử dụng một trong các phương pháp sau:

+ Phương pháp MMPBSA hoặc MMGBSA. Đây là phương pháp nhanh nhất nhưng có độ chính xác không cao và không quá tệ về mặt sai số.

+ Phương pháp metadynamics, umbrella sampling, free energy perturbation (FEP), steered MD … Đây là các phương pháp cao cấp đưa ra sai số nhỏ dựa trên rất nhiều mô phỏng vì vậy cần rất nhiều thời gian và cấu hình máy phải cao.

+ Phương pháp học máy (machine learning): đây là phương pháp khá mới đang được phát triển và hoàn thiện.

Để trả lời cho bài toán đã đặt ra thì ở bài tiểu luận này, em sẽ tìm hiểu về phương pháp học máy (machine learning).

Phương pháp này được tham khảo từ bài báo quốc tế *Bioinfomatics 2018*, tập 34, số 21. Với hướng đi là dựa trên *database PDBbind*, *pafnucy*[1].

I. Khái quát.

[1] https://gitlab.com/cheminfIBB/pafnucy

Kỹ thuật sàng lọc ảo dựa rên cấu trúc là một trong những phương pháp thành công nhất để tăng cường quá trình khám phá thuốc. Với sàng lọc dựa trên cấu trúc, chúng ta sẽ cố gắng dự đoán năng lượng liên kết hoặc những số liệu liên quan giữa phân tử và mục tiêu dựa trên cấu trúc 3D của phức hợp của chúng. Điều này cho phép ta xếp hạng và ưu tiên các phân tử để xử lý thêm và tiến hành các thử nghiệm tiếp theo. Nhiều sơ đồ tính điểm đã được phát triển để hỗ trợ quá trình này. Hiện tại, các phương pháp tính điểm này có một số thay đổi đáng chú ý nhờ dựa trên phương pháp *học máy*.

Học máy sử dụng các hàm tùy ý với các tham số có thể điều chỉnh có khả năng chuyển đổi đầu vào (phức hợp phối tử protein) thành đầu ra (lực liên kết protein - phối tử). Tóm lại, khi mô hình được trình bày với các ví dụ về dữ liệu đầu vào được ghép nối với kết quả mong muốn, nó sẽ '*học*' cách trả về các dự đoán phù hợp với các giá trị được cung cấp. Thông thường quá trình học tập là tăng dần; bằng cách đưa ra những thay đổi nhỏ đối với các tham số của mô hình, dự đoán sẽ được di chuyển đến gần giá trị mục tiêu hơn.

Tuy nhiên phương pháp này có nhược điểm là chúng vẫn dựa vào kỹ thuật tính năng, tức là chúng sử dụng kiến thức chuyên môn để xác định các quy tắc sẽ trở thành cơ sở cho quá trình tiền xử lý dữ liệu đầu vào. Do đó, người ta có thể lập luận rằng chúng chỉ là những hàm tính điểm cổ điển phức tạp hơn với những quy tắc phức tạp hơn. Để giảm thiểu hạn chế, các nhà nghiên cứu đã nghiên cứu phương pháp *học sâu* (deep learning): mô hình học cách trích xuất các tính năng như một kết quả tự nhiên của quá trình khớp các tham số của mô hình với dữ liệu có sẵn; qua đó có thể kết hợp vào mô hình *học máy*.

Các nhà nghiên cứu đã phát triển *pafnucy* một mạng lưới thần kinh sâu mới được thiết kế riêng cho nhiều phương pháp tiếp cận dựa trên cấu trúc, bao gồm ưu tiên phái sinh và sàng lọc ảo. Cấu trúc đầu vào được biểu diễn bằng lưới 3D và sử dụng sự kết hợp giữa các lớp chập và dày đặc; tuy nhiên, mô hình này cố gắng dự

đoán giá trị ái lực ràng buộc chính xác. Pafnucy sử dụng cách tiếp cận tự nhiên hơn để mô tả nguyên tử trong đó cả protein và phối tử đều có cùng loại nguyên tử. Cách tiếp cận này đóng vai trò như một kỹ thuật chính quy hóa vì nó buộc mạng phải khám phá các đặc tính chung của sự tương tác giữa protein và phối tử. Mạng được triển khai bởi *TensorFlow*[1] và được đào tạo dựa trên cơ sở dữ liệu của *PDBbind.*[2]

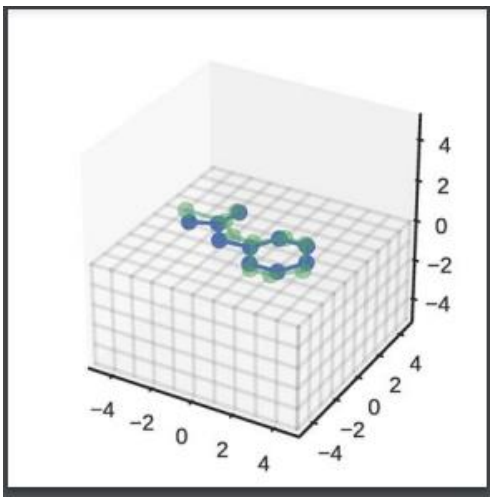
II. *Pafnucy.*

1. Dữ liệu.

1.1. Biểu diễn phức hợp phân tử.

Cấu trúc ba chiều của phức hợp phối tử protein đòi hỏi các phép biến đổi và mã hóa cụ thể để được mạng lưới thần kinh sử dụng. Theo cách tiếp cận của mình, các nhà nghiên cứu đã cắt tổ hợp thành kích thước xác định là hộp 20-Å khối tập trung ở tâm hình học của phối tử. Sau đó, chúng tôi rời rạc hóa vị trí của các nguyên tử nặng bằng cách sử dụng lưới 3D có độ phân giải 1-Å.

[1] Abadi M. et al. (2015) TensorFlow: large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv: 1603.04467.*
[2] Liu Z. et al. (2017) Forging the basis for developing protein–ligand interaction scoring functions. *Accounts Chem. Res.*, 50, 302–309.



Hình 1.1: rời rạc hóa vị trí của các nguyên tử nặng bằng cách sử dụng lưới 3D có độ phân giải 1-Å

Cách tiếp cận này cho phép biểu diễn đầu vào dưới dạng một tensor 4D trong đó mỗi điểm được xác định bởi tọa độ Descartes (3 chiều đầu tiên của tensor) và một vectơ đặc trưng (chiều cuối cùng).

Trong Pafnucy, 19 đặc điểm được sử dụng để mô tả một nguyên tử:

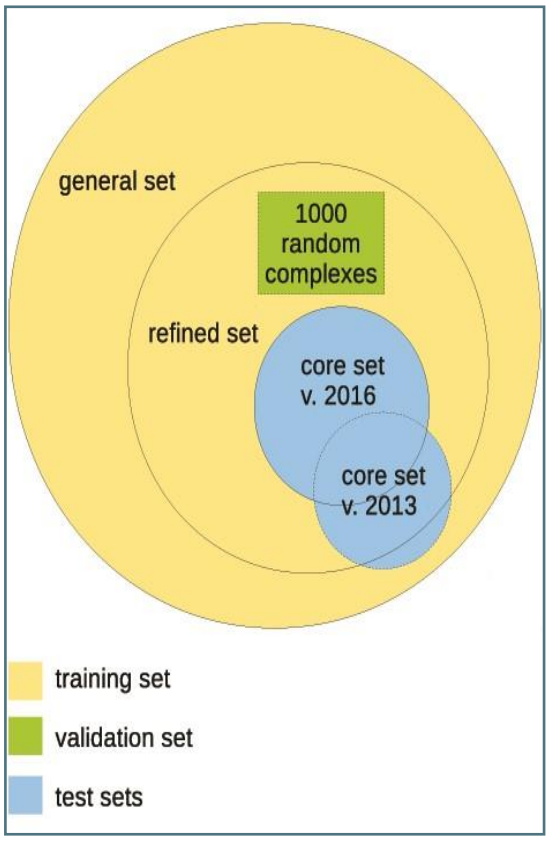
- Các loại nguyên tử mã hóa 9 bit (một nóng hoặc tất cả null): *B*, *C*, *N*, *O*, *P*, *S*, *Se*, *halogen* và *kim loại*
- 1 số nguyên (1, 2 hoặc 3) với atom hybridization: *hyb*
- 1 số nguyên đếm số lượng liên kết với các nguyên tử nặng khác: *Heavy_valence*
- 1 số nguyên đếm số lượng liên kết với các nguyên tử khác loại khác: *Hetero_valence*
- Thuộc tính mã hóa 5 bit (1 nếu có) được xác định bằng mẫu SMARTS: *hydrophobic*, *aromatic*, *acceptor*, *donor* and *ring*
- 1 phao tích điện một phần: *tích điện một phần*
- 1 số nguyên (1 cho phối tử, -1 cho protein) để phân biệt giữa hai phân tử: *moltype*

1.2. Bộ dữ liệu.

Mạng đã được huấn luyện và thử nghiệm với các phức hợp phối tử protein từ cơ sở dữ liệu PDBbind.[1] Cơ sở dữ liệu này bao gồm các cấu trúc 3D của các phức hợp phân tử và ái lực liên kết tương ứng của chúng được biểu thị bằng giá trị pK_a ($-\log K_d$ hoặc $-\log K_i$).

Các phức hợp PDBBind được phân chia thành 3 tập con chồng lên nhau:
Tập chung (general set).
Tập tinh chỉnh (refined set).
Tập lõi (core set).

Các *tập chung* và *tập tinh chỉnh* được sử dụng để huấn luyện mô hình và chọn các siêu tham số, trong khi *tập lõi* được sử dụng làm tập kiểm tra bên ngoài mà mô hình chưa biết trong quá trình huấn luyện và xác nhận. Các tổ hợp còn lại của tập dữ liệu PDBbind được chia như sau: (i) 1000 tổ hợp được chọn ngẫu nhiên từ tập tinh chỉnh đã được sử dụng để xác thực, (ii) toàn bộ tập hợp lõi (290 tổ hợp) được sử dụng làm bộ kiểm tra bên ngoài, (iii) tất cả các tổ hợp khác (phần còn lại của tập tinh chỉnh và tập chung, tổng cộng là 11906) được sử dụng làm tập huấn luyện.



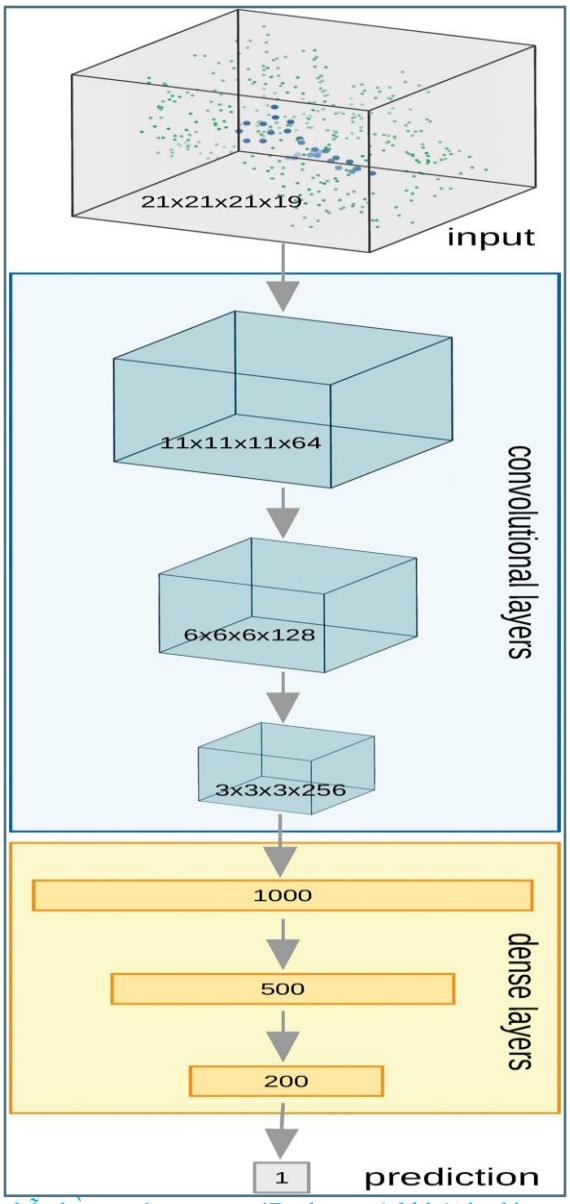
Hình 1.2: Sơ đồ minh họa mối quan hệ giữa các tập hợp con và phân vùng tập dữ liệu có sẵn trong

[1] Liu Z. et al. (2017) Forging the basis for developing protein–ligand interaction scoring functions. *Accounts Chem. Res.*, 50, 302–309.

2. Mạng và phương pháp đào tạo.

2.1. Cấu trúc mạng.

Cấu trúc được sử dụng trong Pafnucy là một mạng lưới thần kinh tích chập sâu với một nơron đầu ra duy nhất để dự đoán mối quan hệ ràng buộc. Mô hình bao gồm hai phần: phần tích chập và phần dày đặc, với các loại kết nối khác nhau giữa các lớp *Tích chập*, từ đó có tên *'tích chập'*, là một phép toán kết hợp hai hàm lại với nhau. Hầu hết các thư viện mạng thần kinh thực sự thay thế hoạt động tích chập bằng tương quan chéo, có cách giải thích trực quan hơn và đo lường sự giống nhau của hai hàm. Mô hình phát hiện các mẫu được mã hóa bởi các bộ lọc trong lớp chập và tạo bản đồ đặc trưng với các lần xuất hiện trong không gian cho từng mẫu trong dữ liệu.



Hình 2.1: Kiến trúc của Pafnucy. Phức hợp phân tử được biểu diễn bằng một tensor 4D, được xử lý bởi ba lớp chập và ba lớp dày đặc (được kết nối đầy đủ) để dự đoán ái lực liên kết

Đầu vào của Pafnucy—phức hợp phân tử—được biểu diễn bằng một tensor 4D và được xử lý giống như một hình ảnh 3D với nhiều kênh màu. Mỗi vị trí của đầu vào (tọa độ x, y và z) được mô tả bằng một vectơ gồm 19 thuộc tính, tương tự như cách mỗi pixel của một hình ảnh (tọa độ x và y) được mô tả bởi một vectơ cường độ của ba màu cơ bản.

Đầu tiên, đầu vào được xử lý bởi một khối gồm các lớp chập 3D kết hợp với lớp gộp tối đa. Pafnucy sử dụng 3 lớp tích chập với 64, 128 và 256 bộ lọc. Mỗi

lớp có các bộ lọc khối 5Å và theo sau là lớp tổng hợp tối đa với một miếng vá khối 2Å.

Kết quả của lớp chập cuối cùng được làm phẳng và được sử dụng làm đầu vào cho một khối các lớp dày đặc (được kết nối đầy đủ). Các nhà nghiên cứ đã sử dụng 3 lớp dày đặc với 1000, 500 và 200 nơ-ron. Để cải thiện tính tổng quát hóa, bỏ học với xác suất bỏ học là 0,5 đã được sử dụng cho tất cả các lớp dày đặc. Họ cũng đã thử nghiệm với 0,2 lần bỏ học và không bỏ học và đạt được kết quả tệ hơn trên bộ xác thực.

Cả hai lớp tích chập và dày đặc đều bao gồm các đơn vị tuyến tính được chỉnh lưu (ReLU). ReLU được chọn vì nó tăng tốc quá trình học tập so với các loại kích hoạt khác.

2.2. Đào tạo.

Các giá trị ban đầu của trọng số bộ lọc tích chập được rút ra từ phân phối chuẩn bị cắt cụt với giá trị trung bình là 0 và độ lệch chuẩn 0,001 và độ lệch tương ứng được đặt thành 0,1. Các trọng số trong các lớp dày đặc được khởi tạo với phân phối chuẩn bị cắt ngắn với giá trị trung bình là 0 và độ lệch chuẩn là $1/\sqrt{n}$, trong đó *n* là số lượng nơ-ron đến cho một lớp nhất định. Độ lệch tương ứng được đặt thành 1.0.

Trình tối ưu hóa Adam được sử dụng để huấn luyện mạng với 10^{-5} tốc độ học tập và 5 ví dụ cho mini batch (Bộ huấn luyện chứa 11906 tổ hợp; do đó, lô cuối cùng thực sự bao gồm 6 tổ hợp thay vì 5.). Với batch size lớn hơn (10 đến 20 ví dụ) cũng đã được thử nghiệm nhưng cho hiệu suất kém hơn. Quá trình đào tạo được thực hiện trong 20 epochs và mô hình có sai số thấp nhất trên bộ xác thực đã được chọn (trong trường hợp mạng được mô tả trong công việc này là sau 14 epochs đào tạo).

Để giảm việc trang bị quá mức, họ đã sử dụng phương pháp bỏ học được đề cập trước đó và giảm trọng lượng L2 với $\lambda=0,001$. Sử dụng giá trị cao hơn

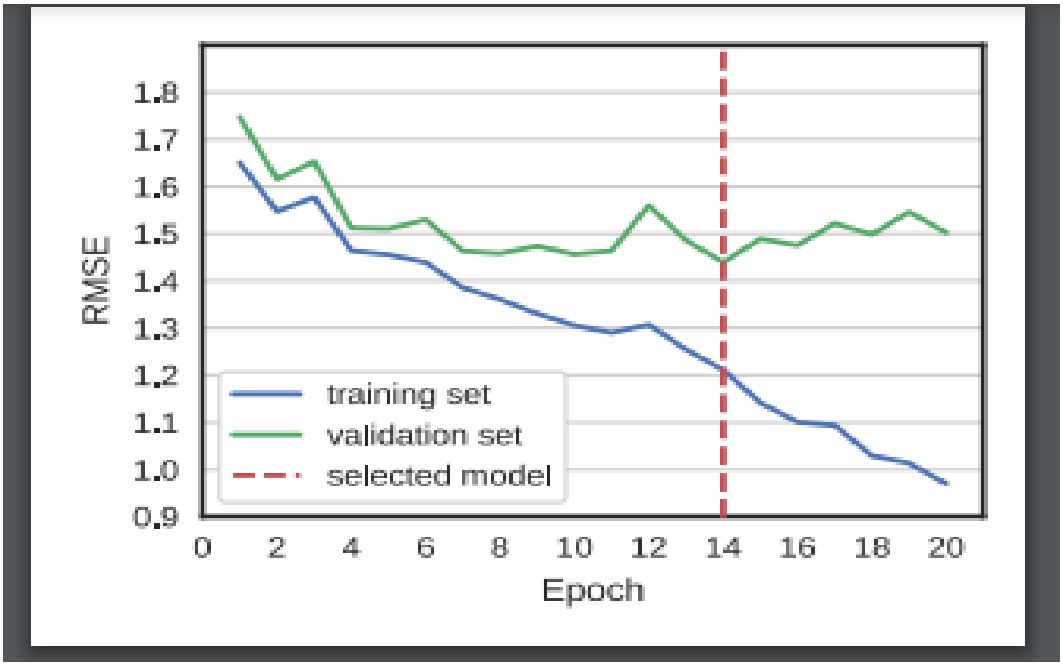
(λ=0,01) đã giảm công suất của mô hình quá nhiều và dẫn đến lỗi xác thực và đào tạo cao hơn. Ngoài việc cung cấp tính chính quy, L2 cho phép họ điều tra tầm quan trọng của tính năng. Nếu trọng số khác 0 đáng kể, thông tin mà nó truyền tải phải quan trọng để mô hình đưa ra dự đoán.

Một phần quan trọng trong cách tiếp cận này là phát triển một mô hình không nhạy cảm với định hướng phức tạp của thụ thể phối tử. Do đó, mọi cấu trúc được đưa vào mạng theo 24 hướng khác nhau (tức là tất cả các kết hợp có thể có của 90°quay của một hộp lập phương), thu được 24 ví dụ huấn luyện khác nhau cho mỗi phức hợp phối tử protein.

Bằng cách sử dụng các vòng quay phức tạp có hệ thống trong quá trình đào tạo, họ dự đoán rằng mạng sẽ tìm hiểu các quy tắc chung hơn về tương tác phối tử protein và dẫn đến hiệu suất tốt hơn trên dữ liệu mới. Thật vậy, trong các thử nghiệm của mình, họ đã quan sát thấy hiệu suất kém hơn nhiều của các mô hình được đào tạo theo các hướng duy nhất bất kể siêu tham số được sử dụng để xác định một mạng cụ thể.

3. Kết quả.

Lỗi trên tập huấn luyện và xác nhận đã được theo dõi trong quá trình học



Hình 3.1: Lỗi trên tập huấn luyện và xác nhận đã được theo dõi trong quá trình học

Mặc dù mô hình đã được huấn luyện trên 24 phép quay khác nhau của mỗi phức, *RMSE* (sai số bình phương trung bình gốc) chỉ được tính cho hướng ban đầu để tăng tốc độ tính toán.

Sau 14 epochs đào tạo, mô hình bắt đầu quá khớp và lỗi trên bộ xác thực bắt đầu tăng chậm nhưng đều đặn. Tập trọng số tốt nhất của mạng thu được sau 14 epochs đào tạo đã được lưu và sử dụng làm mô hình cuối cùng. Hiệu suất của mô hình được đánh giá trên tất cả các tập hợp con của dữ liệu.

Bảng 1: hiệu xuất của pafnucy

Tập dữ liệu	**RMSE**	**MAE**	**SD**	**R**
Kiểm tra	1.42	1.13	1.37	0.78
Thẩm định	1.44	1.14	1.43	0.72
Đào tạo	1.21	0.95	1.19	0.77

Đối với mỗi phức hợp trong tập dữ liệu, mối quan hệ được dự đoán và so sánh với giá trị thực. Lỗi dự đoán được đo bằng *RMSE* và *MAE* (sai số tuyệt đối trung

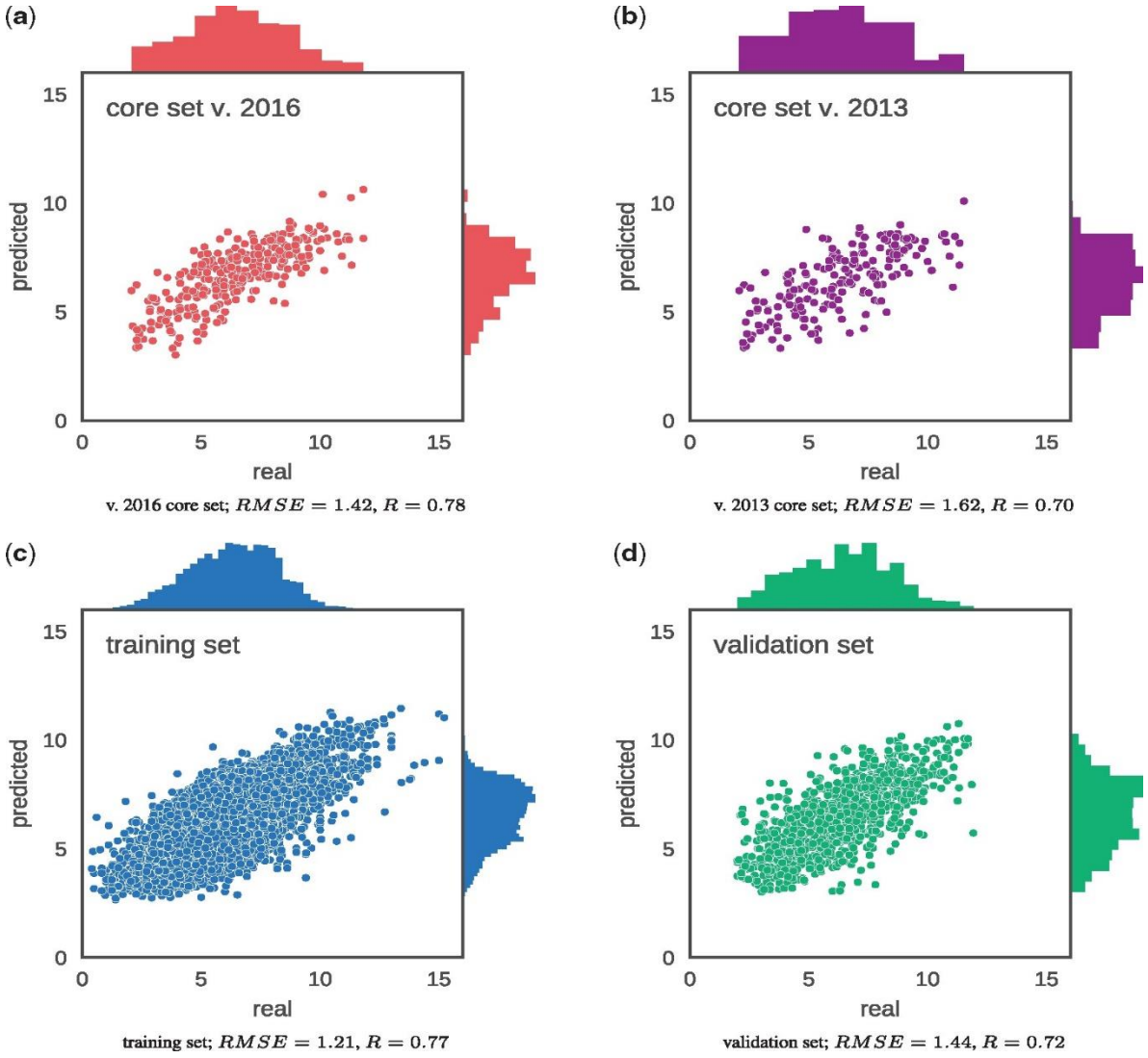
bình). Mối tương quan giữa điểm số và hằng số liên kết được đo bằng thực nghiệm được đánh giá bằng hệ số tương quan Pearson (R) và độ lệch chuẩn trong hồi quy (SD). SD là thước đo được sử dụng trong CASF[1] và được định nghĩa như sau:

$$\mathrm{SD} = \sqrt{\frac{1}{\mathrm{N}-1}\sum_{\mathrm{i}=1}^{\mathrm{N}}\{\mathrm{ti} - (\mathrm{ayi}+\mathrm{b})\}^2}$$

trong đó t_i và y_i là các mối quan hệ được đo và dự đoán cho phức thứ i , trong khi a và b lần lượt là độ dốc và điểm giao nhau của đường hồi quy giữa các giá trị được đo và dự đoán.

Dự đoán cho hai bộ kiểm tra (bộ lõi từ PDBbind v. 2016 và v. 2013), bộ huấn luyện và bộ xác thực

[1] Li Y. et al. (2014) Comparative assessment of scoring functions on an updated benchmark: 2. evaluation methods and general results. *J. Chem. Inf. Model*., 54, 1717–1736.



Hình 3.2: Dự đoán cho hai bộ kiểm tra (bộ lõi từ PDBbind v. 2016 và v. 2013), bộ huấn luyện và bộ xác thực

Đúng như kỳ vọng, mạng đạt được sai số thấp nhất trên tập huấn luyện (*Hình 3.2c*), tập này được sử dụng để tìm trọng số của mạng. Quan trọng hơn, Pafnucy cũng trả về các dự đoán chính xác cho hai bộ thử nghiệm (*Hình 3.2a và b*), mà mô hình chưa biết trong quá trình đào tạo và xác nhận.

Các kết quả trên điểm chuẩn *scoring power* CASF-2013 (bộ lõi PDBbind v. 2013), mặc dù kém hơn đáng kể so với các tập hợp con khác, nhưng vẫn tốt hơn

so với kết quả của bất kỳ chức năng tính điểm nào khác[1]: X-Score hoạt động tốt nhất có $R = 0{,}61$ và $SD = 1{,}78$, trong khi pafnucy đạt được $R = 0{,}70$ và $SD = 1{,}61$. Mô hình duy nhất có hiệu suất tốt hơn được công bố cho đến nay là RF-Score v3, đạt $R = 0{,}74$ và $SD = 1{,}51$ trên CASF-2013.[2]

	Pafnucy	**X-Score**	**ChemScore**	**ChemPLP**	**PLP1**	**G-Score**
SD	1.61	1.78	1.82	1.84	1.86	1.87
R	0.7	0.61	0.59	0.58	0.57	0.56

Bảng 2: kết quả trên điểm chuẩn của CASF 2013

Đồng thời cũng so sánh pafnucy với X-Score trên *Astex Diverse set.*

Bảng 3: Độ chính xác của dự đoán trên *Astex Diveser set*

Phương pháp	**RMSE**	**MAE**	**SD**	**R**
Pafnucy	1.43	1.13	1.43	0.57
X-Score	1.55	1.22	1.48	0.52

Cả hai phương pháp đều có lỗi tương đương với lỗi thu được trên dữ liệu PDBbind. Đúng như dự đoán, Pafnucy vượt trội hơn X-Score trên *Astex Diverse set*, bất kể sử dụng thước đo nào. Tuy nhiên, mối tương quan quan sát được thấp hơn đối với cả hai phương pháp. Hiệu ứng này một phần là do tập dữ liệu Astex

[1] Li Y. et al. (2014) Comparative assessment of scoring functions on an updated benchmark: 2. evaluation methods and general results. *J. Chem. Inf. Model*., 54, 1717–1736.
[2] WójcikowskiM. et al. (2015) Open drug discovery toolkit (oddt): a new open-source player in the drug discovery field. *J. Cheminf*., 7, 26

chỉ chứa 73 phức hợp và do đó, mối tương quan nhạy cảm hơn nhiều với những thay đổi nhỏ trong dự đoán so với các tập hợp con lớn hơn.

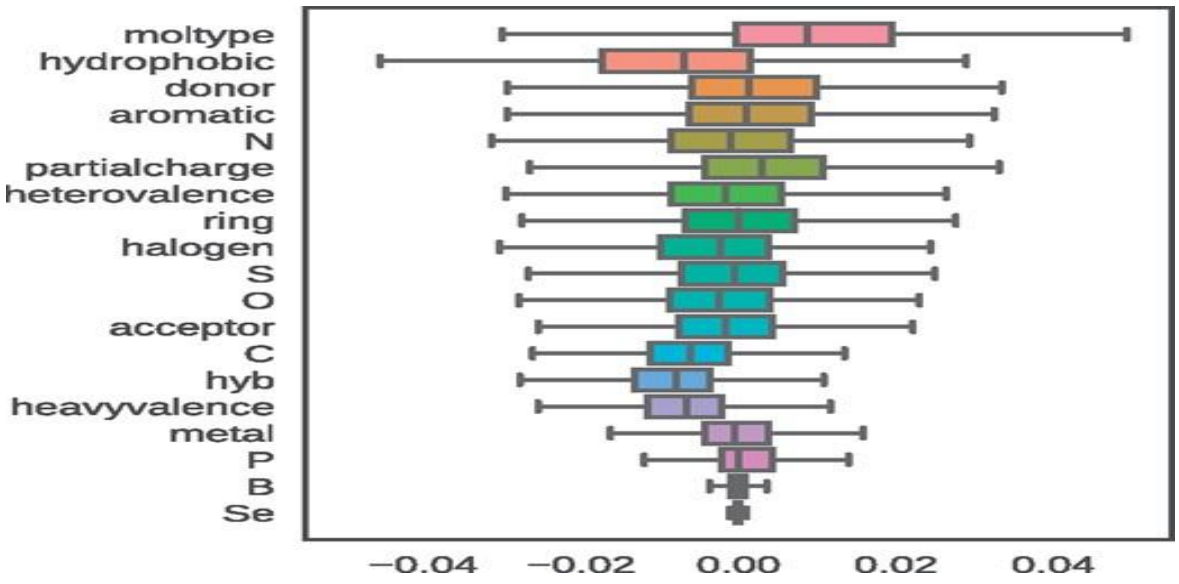
4. Thảo luận.

4.1. Cách Pafnucy xem và xử lý.

Mạng lưới thần kinh thường được cho là khó phân tích và diễn giải hơn các mô hình đơn giản khác. Điều đáng lo ngại là một mô hình có thể mang lại những dự đoán tốt vì những lý do sai lầm (ví dụ: các thành phần ẩn trong dữ liệu) và do đó sẽ không khái quát hóa tốt cho các tập dữ liệu mới. Để tin tưởng vào mạng lưới thần kinh và các dự đoán của nó, người ta cần đảm bảo rằng mô hình sử dụng thông tin có liên quan đến nhiệm vụ hiện tại. Vì vậy việc phân tích dữ liệu đầu vào là quan trọng nhất và có tác động lớn nhất đến khả năng dự đoán của mô hình.

Với mô hình Pafnucy, được đào tạo với L2 (với λ=0,001), chúng ta có thể ước tính tầm quan trọng của đặc điểm bằng cách xem xét sự phân bổ trọng số liên quan đến các bộ lọc tích chập trong lớp ẩn đầu tiên. Giá trị ban đầu của chúng gần bằng 0. Trong quá trình học tập, các trọng số có xu hướng lan rộng và hình thành phạm vi rộng hơn, vì các trọng số có giá trị tuyệt đối cao hơn sẽ truyền nhiều thông tin hơn đến các lớp sâu hơn của mạng. Bởi vì Pafnucy đã được học tập với chính quy hóa L2 nên chỉ có các trọng số quan trọng mới có giá trị tuyệt đối cao như vậy.

Đầu vào của Pafnucy được biểu diễn bằng 19 đặc điểm.

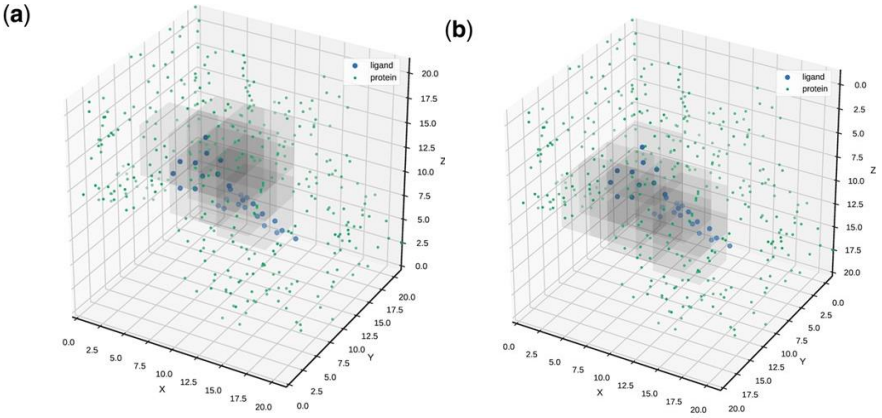


Hình 4.1 Phạm vi trọng số cho từng kênh đầu vào (tính năng). Các ngoại lệ không được hiển thị

Ở *hình 4.1*, ta có thể thấy được đặc điểm có phạm vi rộng nhất là đặc điểm *moltype* - đặc điểm phân biệt protein với phối tử. Kết quả này ngụ ý rằng Pafnucy đã học được rằng ái lực liên kết phụ thuộc vào mối quan hệ giữa hai phân tử và việc nhận ra chúng là rất quan trọng. Ngoài ra, trọng số của các loại nguyên tử selen và boron (*Se* và *B* , tương ứng) hầu như không thay đổi trong quá trình học tập và gần như bằng 0. Kết quả này có thể được giải thích theo hai cách: hoặc mạng tìm thấy các tính năng khác của phức hợp phối tử protein quan trọng hơn đối với ái lực liên kết hoặc do sự xuất hiện không thường xuyên của các loại nguyên tử này trong các phối tử nên mạng không thể tìm thấy bất kỳ mô hình chung nào về ảnh hưởng của chúng lên ái lực liên kết.

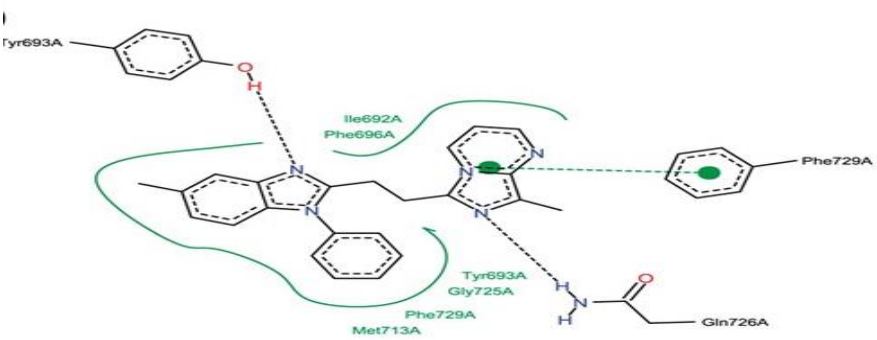
Để kiểm tra sâu hơn về dữ liệu đầu vào, các nhà nghiên cứu đã phân tích tác động của việc thiếu dữ liệu đầu vào đối với khả năng dự đoán. Để kiểm tra điều này, họ đã chọn một trong các phức PDE10A có chất ức chế benzimidazole. Thí nghiệm được thực hiện như sau: 343 phức hợp bị hỏng được tạo ra với một số dữ liệu bị thiếu và dự đoán ái lực liên kết cho từng phức hợp. Dữ liệu bị thiếu được tạo ra bằng cách xóa hộp 5-Å khỏi dữ liệu gốc, trượt hộp với bước 3-Å (theo mọi hướng), do đó thu được7^3=343 đầu vào bị hỏng. Tiếp theo, xoay phức hợp 180°

về trục X và làm theo quy trình tương tự, do đó tạo ra thêm 343 đầu vào bị hỏng. Sau đó, đối với mỗi hướng trong số hai hướng, lấy 10 đầu vào bị hỏng có mức giảm ái lực dự đoán cao nhất.



Hình 4.2: Các phần quan trọng nhất của đầu vào. Bất kể hướng phức tạp như thế nào, cùng một vùng đầu vào có tác động cao nhất đến dự đoán. Lưu ý rằng đồ thị thứ hai được quay lại quanh trục X để dễ so sánh. (**a**) Định hướng ban đầu. (**b**) Quay bởi 180°về trục X.

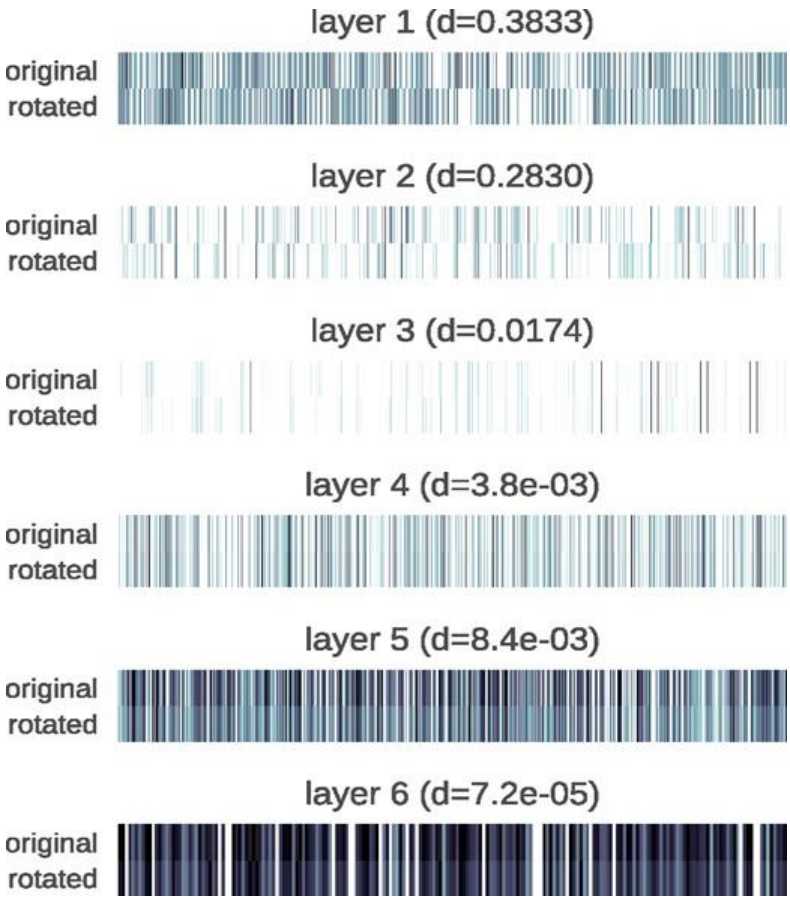
Như đã thấy ở *hình 4.2a* (hướng ban đầu) và *hình 4.2b* (quay góc 180^{o} theo trục X), đối với cả hai hướng, ta đã xác định cùng một vùng chứa phối tử và vùng lân cận gần nhất của nó.



Hình 4.3: Tương tác phối tử protein. Đồ họa được tạo bằng Poseview (Stierand và Rarey, 2010)

Ở *hình 4,3*: Các hộp chứa các amino acids tham gia tương tác với phối tử, tức là Gln726, tạo thành liên kết hydro và Phe729, tạo thành liên kết π–π tương tác với phối tử.

Quay trở lại với dữ liệu đầu vào không bị hỏng, các nhà nghiên cứu muốn nghiên cứu cách Pafnucy quản lý để đưa ra các dự đoán gần như giống hệt nhau cho hai hướng khác nhau của phức (hướng thứ hai quay quanh trục X bằng180°). Đối với yêu cầu này, họ đã phân tích việc kích hoạt các lớp ẩn cho hai đầu vào.



Hình 4.4: Kích hoạt trên các lớp ẩn cho hai hướng của phức hợp PDE10A (ID PDB: 3WS8). Màu tối hơn cho thấy giá trị cao hơn. Khoảng cách cosine (d) giữa các mẫu kích hoạt cho mỗi lớp được cung cấp

Ở *hình 4.4:* chúng ta có thể thấy lớp ẩn đầu tiên có các mẫu kích hoạt rất khác nhau cho hai hướng của đầu vào. Pafnucy nhận được dữ liệu rất khác nhau và cần sử dụng các bộ lọc khác nhau trong lớp tích chập đầu tiên để xử lý chúng. Tuy nhiên, chúng ta càng đến gần lớp đầu ra thì các kích hoạt càng giống nhau. Chúng ta có thể thấy rõ rằng mô hình của chúng ta đã học cách trích xuất cùng một thông tin từ dữ liệu được trình bày khác nhau.

III. Kết luận.

Trong nghiên cứu này, chúng tôi đã trình bày một mạng lưới thần kinh sâu – Pafnucy - có thể được sử dụng trong các chiến dịch khám phá phối tử dựa trên cấu trúc; như một chức năng tính điểm trong sàng lọc ảo hoặc dự đoán ái lực đối với các phân tử mới sau khi phức hợp được tạo ra. Pafnucy cũng có thể được sử dụng trực tiếp trong quá trình lắp ghép để hướng dẫn tối ưu hóa tư thế phối tử. Mô hình đã được thử nghiệm trên tiêu chuẩn *Scoring power* CASF-2013 và vượt trội hơn tất cả 20 mô hình tính điểm hiện đại được các nhà nghiên cứu CASF-2013 thử nghiệm. Kết quả thu được và phân tích cẩn thận về mạng cho thấy Pafnucy đưa ra những dự đoán đáng tin cậy dựa trên các tính năng có liên quan.

Dự đoán tác động của các phân tử nhỏ lên các mục tiêu protein quan trọng về mặt sinh học đa dạng đã được các nhà nghiên cứu tìm kiếm từ lâu. Pafnucy có thể được áp dụng để kiểm tra nhiều hợp chất đối với một protein hoặc để kiểm tra nhiều protein đối với một hợp chất. Do đó, nó có thể giúp khám phá các loại thuốc tiềm năng mới cũng như nghiên cứu tác dụng phụ của các phân tử hoạt tính sinh học. Bằng cách dự đoán tác động tiềm tàng của các loại thuốc mới đối với sinh học của tế bào, Pafnucy có thể đóng góp cho các ngành như y học hệ thống và sinh học hệ thống.

Tài liệu tham khảo.

1. A machine learning approach to predicting protein–ligand binding affinity with applications to molecular docking

Pedro J. Ballester, John B. O. Mitchell
Bioinformatics, Volume 26, Issue 9, May 2010, Pages 1169–1175,

2. Development and evaluation of a deep learning model for protein–ligand binding affinity prediction

Marta M Stepniewska-Dziubinska, Piotr Zielenkiewicz, Pawel Siedlecki
Bioinformatics, Volume 34, Issue 21, November 2018, Pages 3666–3674,

3. Structure-Based Virtual Screening for Drug Discovery: a Problem-Centric Review

Tiejun Cheng, Qingliang Li, Zhigang Zhou, Yanli Wang

4. Forging the Basis for Developing Protein–Ligand Interaction Scoring Functions

Zhihai Liu, Minyi Su[†], Li Han,[†]Jie Liu, [†]Qifan Yang, [†]Yan Li[*] and Renxiao Wang[†‡]*

5. Goodfellow I. et al. (2016)*Deep Learning*, MIT Press, Cambridge, MA

6. Open Drug Discovery Toolkit (ODDT): a new open-source player in the drug discovery field

Maciej Wójcikowski, Piotr Zielenkiewicz & Pawel Siedlecki Journal of Cheminformatics volume 7, Article number: 26 (2015)

7. Drawing the PDB: Protein−Ligand Complexes in Two Dimensions

Katrin Stierand and Matthias Rarey[*]

8. https://gitlab.com/cheminfIBB/pafnucy