



Predicting Molecular Docking of Per- and Polyfluoroalkyl Substances to Blood Protein Using Generative Artificial Intelligence Algorithm Diffdock

Dhan Lord B Fortela, Ashley P Mikolajczyk, Miranda R Carnes, Wayne Sharp, Emmanuel Revellame, Rafael Hernandez, William E Holmes & Mark E Zappi

To cite this article: Dhan Lord B Fortela, Ashley P Mikolajczyk, Miranda R Carnes, Wayne Sharp, Emmanuel Revellame, Rafael Hernandez, William E Holmes & Mark E Zappi (2024) Predicting Molecular Docking of Per- and Polyfluoroalkyl Substances to Blood Protein Using Generative Artificial Intelligence Algorithm Diffdock, BioTechniques, 76:1, 14-26, DOI: [10.2144/btn-2023-0070](https://doi.org/10.2144/btn-2023-0070)

To link to this article: <https://doi.org/10.2144/btn-2023-0070>



© 2024 The Authors



[View supplementary material](#)



Published online: 10 Nov 2023.



[Submit your article to this journal](#)



Article views: 1698



[View related articles](#)



[View Crossmark data](#)

Predicting molecular docking of per- and polyfluoroalkyl substances to blood protein using generative artificial intelligence algorithm DiffDock

Dhan Lord B Fortela^{*,1,2} , Ashley P Mikolajczyk^{1,2}, Miranda R Carnes¹, Wayne Sharp^{2,3}, Emmanuel Revellame^{1,2} , Rafael Hernandez^{1,2} , William E Holmes^{1,2}  & Mark E Zappi^{1,2} 

¹Department of Chemical Engineering, University of Louisiana, Lafayette, LA 70504, USA; ²Energy Institute of Louisiana, University of Louisiana, Lafayette, LA 70504, USA; ³Department of Civil Engineering, University of Louisiana, Lafayette, LA 70504, USA; *Author for correspondence: dhanlord.fortela@louisiana.edu

BioTechniques 76: 15–26 (January 2024) 10.2144/btn-2023-0070

First draft submitted: 9 August 2023; Accepted for publication: 27 October 2023; Published online: 10 November 2023

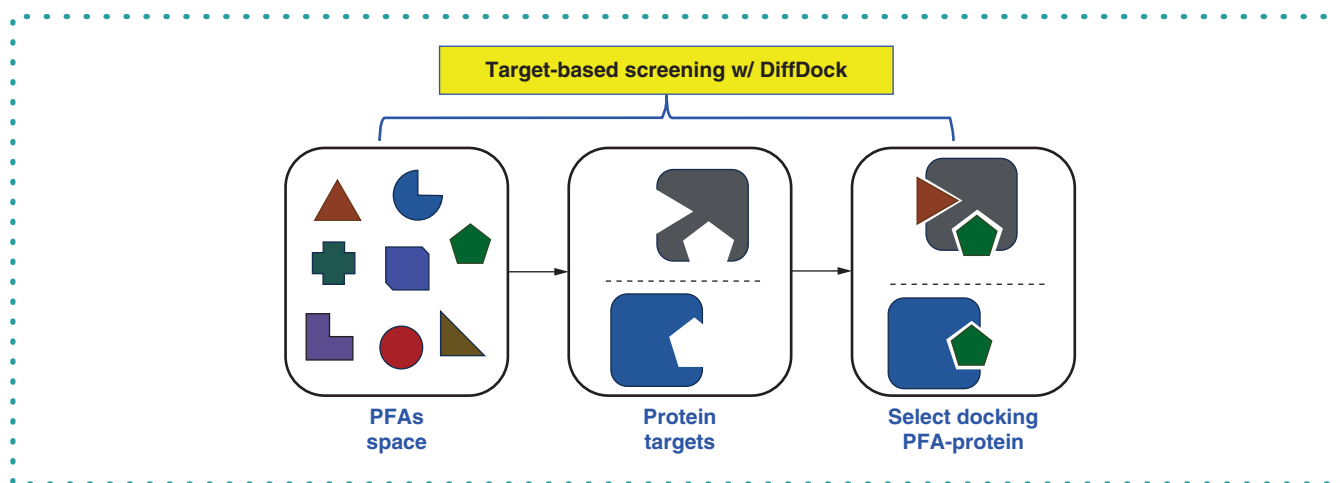
ABSTRACT

This study computationally evaluates the molecular docking affinity of various perfluoroalkyl and polyfluoroalkyl substances (PFAs) towards blood proteins using a generative machine-learning algorithm, DiffDock, specialized in protein–ligand blind-docking learning and prediction. Concerns about the chemical pathways and accumulation of PFAs in the environment and eventually in the human body has been rising due to empirical findings that levels of PFAs in human blood has been rising. DiffDock may offer a fast approach in determining the fate and potential molecular pathways of PFAs in human body.

TWEETABLE ABSTRACT

This study demonstrates the capability of generative AI algorithm DiffDock to accelerate protein PFA molecular docking computations that can lead to efficient studies of PFA fate in the human body.

GRAPHICAL ABSTRACT



KEYWORDS:

blood proteins • generative artificial intelligence • human health • molecular docking • per- and polyfluoroalkyl substances • target-based screening

The risks to human health and environment of perfluoroalkyl and polyfluoroalkyl substances (PFAs) have been a major concern in the current decade due to their prevalence in water, soil, air and food [1,2]. PFAs have been used in industrial manufacturing and consumer products due to their useful properties [3]. Current scientific research suggests that exposure to certain PFAs may be harmful to human health [4,5]. This growing concern about the environmental and anthropologic pathways of PFAs has prompted health and environmental

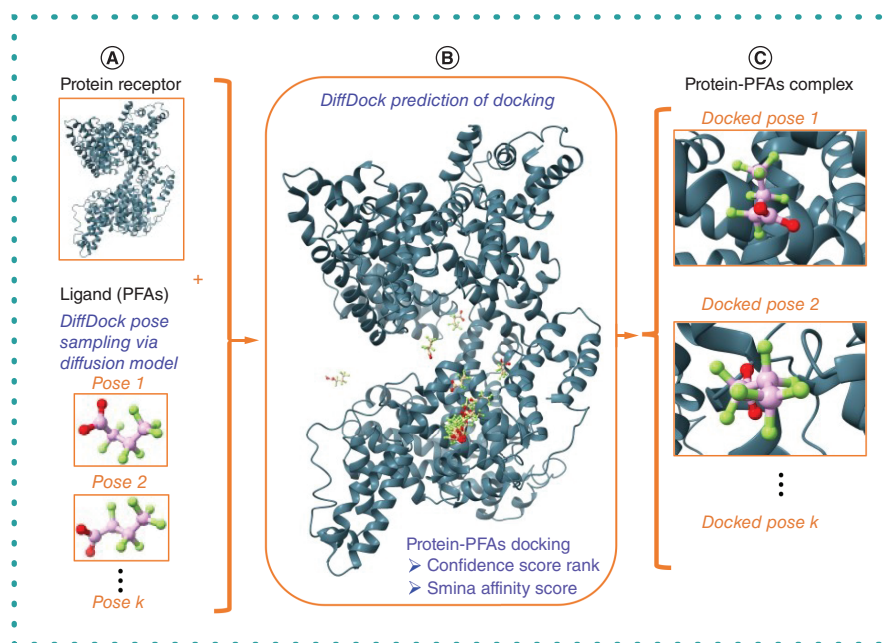


Figure 1. Schematic overview of the data analytics workflow implemented in this study. The example docking shown is that of protein receptor albumin (1AO6) and ligand perfluorobutanoic acid (PFBA). (A) The conformation of the ligand molecule is sampled via diffusion generative model in DiffDock resulting in a ligand pose (with all information about atomic positions in the molecule) to be docked to the protein receptor. (B) The DiffDock algorithm takes the ligand pose and docks the ligand to the protein by searching for the docking position of higher likelihood with 100 inference steps using the probability distribution developed in the trained DiffDock model. The convergence of PFBA to a certain location in the albumin protein is apparent in (B). The associated confidence score and smina affinity score of the docking is also computed. Then another ligand pose is generated, and the docking computations are repeated until all target ligand poses are docked. (C) The confidence score of docking by each ligand pose is then used to rank the ligand poses, and the smina affinity scores are reported. A video clip of the docking inference steps of PFBA-albumin to convergence of Rank 1 pose is provided as Supplementary Materials and in the GitHub repository of the paper.

regulatory agencies such as the US NIH [6] and US EPA to implement strategic plans to address PFA contamination and accumulation [1]. The US CDC has recently established blood-testing procedures for PFAs as part of the efforts to understand the health effects of PFAs [7]. A chemical analysis of human blood plasma indicates that albumin is the major carrier protein for PFAs [8]. Although such blood chemistry analyses are currently being developed and tested via case studies [9], human blood is a complex matrix of biomolecules [10] in which established chemical analysis techniques may not be able to easily detect PFAs complexed with blood components. With human blood serving many functions including the critical task of distributing nutrients to various parts of the body [11], blood-based analyses provides critical insights about human health [12], such as risks with PFAs.

A potential powerful tool that can be used to determine the molecular docking affinity of PFAs with human blood proteins is the generative artificial intelligence (AI) algorithm called DiffDock [13], which is the latest significantly improved generative-learning algorithm for molecular docking of a ligand (or a small molecule) to protein receptors trained on large dataset of protein–ligand complexes [14]. Although originally intended for drug-discovery applications, DiffDock can be applied to other similar protein–ligand docking problems because DiffDock was trained by Corso *et al.* [13] on 17,000 protein–ligand complexes from the Protein Data Bank (PDB) [15]. DiffDock learns over the manifold of ligand poses consists of translational, rotational and torsional dimensions by implementing a diffusion-generative model [13,16].

This work implements the DiffDock algorithm in the prediction of molecular docking of PFAs with blood proteins to illustrate potential uses outside of its original intended application in drug discovery. This work is a first demonstration on the use of generative AI in estimating the affinity of PFAs to bind onto human blood proteins. Even though the work is limited to select human blood proteins, the findings of the work may usher further implementations of a generative AI algorithm such as DiffDock in elucidating molecular binding of PFAs onto other proteins in the human body. In a more general view, the use of generative AI in molecular docking prediction may aid in fast and comprehensive studies of the pathways of PFAs into human bodies and other living organisms, and in the improved design of materials, processes and technologies to minimize or eliminate bioaccumulation of PFAs and other contaminants.

Materials & methods

A schematic of the data analytics workflow implemented is shown in Figure 1.

The DiffDock algorithm computations were implemented using Python coding language [17,18] run in GoogleColab using NVIDIA Tesla T4 GPU runtime [19]. Each run of protein–ligand docking computations took 30 min on average to complete; hence, all runs of

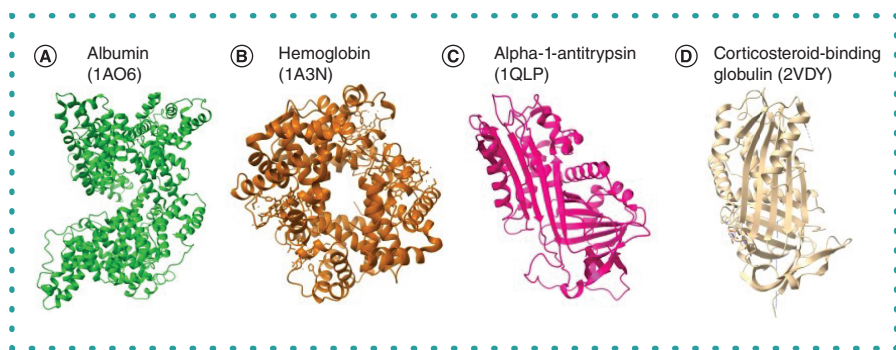


Figure 2. 3D structure of the human blood proteins used in this work. (A) Albumin with Protein Data Bank (PDB) identification number 1AO6. (B) Hemoglobin with PDB identification number 1A3N. (C) Alpha-1-antitrypsin with PDB identification number 1QLP. (D) Corticosteroid-binding globulin with PDB identification number 2VDY.

42 protein–ligand pairs resulted in a total of 1260 min (21 h) of DiffDock runs for the whole dataset used in this paper. A copy of the Python codes used are provided as Jupyter Notebook [20] file in the GitHub repository for the paper [21]. The 3D renderings of protein and molecular structures were carried out using ChimeraX software [22,23]. The structural formula of PFAs were rendered using the OpenBabel software [24].

Study dataset: blood proteins & PFA molecular structures

The human blood proteins used in this analysis were selected to represent key functions in blood serum: 1) albumin, which is 55% of the blood plasma proteins [25]; 2) hemoglobin, which is the transport protein in red blood cells [26]; 3) alpha-1-antitrypsin [27], which is a serpin globulin that is typically concentrated from donated bloods and used for therapy of certain disorders; and 4) corticosteroid-binding globulin (CBG), which is a glycoprotein that binds to cortisol and other glucocorticoids in the blood and helps with anti-inflammatory actions [28]. The blood proteins structures format used was the ‘.pdb’ and the protein molecular data were downloaded from the PDB online repository [15] via the PDB identification label of each protein: 1AO6 for albumin [29,30], 1A3N for hemoglobin [31,32], 1QLP for alpha-1-antitrypsin [33,34] and 2VDY for CBG [35,36]. A 3D rendering of the protein structures are shown in Figure 2.

The PFA molecular structures format used was the simplified molecular-input line-entry system (SMILES) [37] downloaded from PubChem [38], as shown in Supplementary Table 1. Selection of the PFAs was based on the result of a review of the current literature on PFAs with a focus on the reports of the US EPA [1], NIH [6] and published empirical analysis of PFAs in human blood [8,39,40]. Hence, there are 12 PFAs used in the current work ranging from short-chain to long-chain structures (Supplementary Table 1). For a comprehensive analysis, the structural formula and the 3D ball-and-stick rendering of the PFAs are also presented in Figure 3. The PFAs are as follows: perfluorobutanoate (PFB), perfluorobutanoic acid (PFBA), hexafluoropropylene oxide dimer acid (HFPO-DA), perfluorooctanesulfonic acid (PFOS), perfluorooctanoic acid (PFOA), perfluorononanoic acid (PFNA), perfluorohexanesulfonic acid (PFHxS), perfluorobutanesulfonic acid (PFBS), perfluoropentanoic acid (PFPeA), perfluorohexanoic acid (PFHxA), perfluoroheptanoic acid (PFHpA) and perfluorodecanoic acid (PFDA).

DiffDock: docking inference, docking rank & binding affinity score

The DiffDock algorithm implements a generative diffusion model (a.k.a., Brownian motion) during training and inference steps [13]. This generative task allows the efficient sampling of translational, rotational and torsional parameters of a ligand as it docks with the protein receptor [13]. Since DiffDock has been trained on a large dataset of protein–ligand complexes (17,000 protein–ligand complexes from PDB; see Corso *et al.* for details [13]), the pretrained model can be used for inference of protein–ligand complex systems. By default in the DiffDock algorithm, a total 40 docking poses were generated and each ligand pose was docked to the protein using 100 steps. These default settings of ligand–protein docking prediction were used in this work. The goodness-of-fit of protein–ligand docking can be measured using the proposed metrics by the originators of the algorithm [13]: 1) Rank of docking based on Confidence Score and 2) smina Affinity Score. These metrics used in this study are discussed as follows.

Molecular docking rank

The molecular docking pose rank (Equation 1) was based on the confidence score (Equation 2), as proposed by Corso *et al.* [13], which was based on the concept by Song *et al.* [16]. The pertinent equations implemented in the algorithm for this docking rank are as follows. The rank of protein–ligand docking pose is evaluated in DiffDock by sorting the confidence score (Equation 2), where rank 1 is the docking with the highest confidence score.

$$\text{Rank} = \text{sort}(\text{confidence score of docked pose 1}, \dots, \text{pose k}) \quad (\text{Equation 1})$$

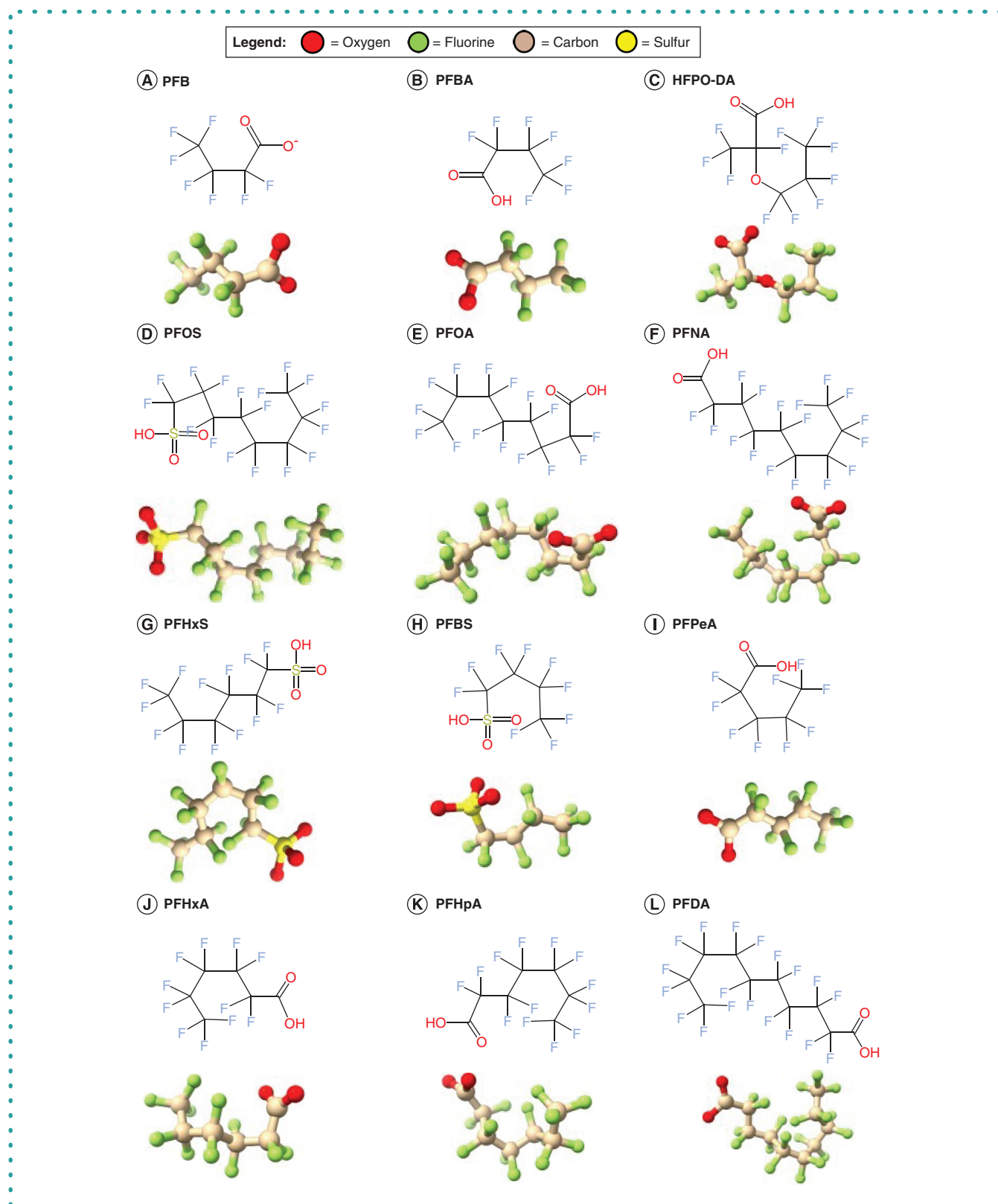


Figure 3. Structural formula and ball-and-stick rendering of the perfluoroalkyl and polyfluoroalkyl substance molecules used in the molecular docking computations with blood proteins. (A) PFB, (B) PFBA, (C) HFPO-DA, (D) PFOS, (E) PFOA, (F) PFNA, (G) PFHxS, (H) PFBS, (I) PFPeA, (J) PFHxA, (K) PFHpA and (L) PFDA.

$$\text{Confidence score of docked pose } x = \nabla_x \log p_t(x) \quad (\text{Equation 2})$$

Equation 2 describes the score of each marginal distribution of ligand pose x , which is the assignment of atomic positions in \mathbb{R}^3 (3D) [13], at each time step, t , of the generative diffusion process [16]. Hence, a ligand with n atoms will have \mathbb{R}^{3n} dimensions in every pose x . The gradient, ∇_x , is the result of framing the diffusion process as stochastic differential equations according to Song *et al.* [16], and p denotes the ligand pose probabilistic distribution. Rigorous derivations of Equations 1 & 2 are presented by Corso *et al.* [13] and Song *et al.* [16].

smina affinity score

The scoring and minimization with AutoDock Vina (smina) affinity score by Koes *et al.* [41] was also integrated for a comprehensive evaluation of the correct docking poses. The general form of the equation implemented in the algorithm for the smina affinity score computations is shown in Equation 3, where c is the smina affinity score (in kcal/mol). The smina score can be positive or negative depending on the strength of the interaction between the protein and the ligand. A negative smina score indicates a stronger binding affinity while a positive smina score indicates weaker binding affinity [41]. The smina affinity score is a measure of the standard chemical potential of the protein–ligand system [42]. Hence, the smina affinity score is an approximation of the Gibbs free energy of binding between the protein receptor and a ligand [43]. The smina affinity score takes the unit of energy per mole, specifically kcal/mol in the DiffDock algorithm [13], which is consistent with AutoDock Vina [42]. The chemical potential of the protein–ligand complex is minimized when the smina affinity score is minimized [42]. Equation 3 is essentially the sum of intermolecular forces c_{inter} and intramolecular forces c_{intra} [42].

$$c = c_{inter} + c_{intra} = \sum_{i < j} f_{t_i t_j}(r_{ij}) \quad (\text{Equation 3})$$

The summation in Equation 3 is over all pairs of atoms that can move relative to each other. The type of atom i is designated by a type t_i and a set of interactions functions $f_{t_i t_j}$ with the interatomic distance r_{ij} [42]. The details of the various components of Equation 3 have been published elsewhere [41–43]. The computation of smina affinity score have been included in the Python codes used to run the DiffDock algorithm and are part of the Jupyter Notebook file in the paper online repository [21].

Results & discussion

The results of DiffDock inferences on the molecular docking of PFAs with blood proteins are presented in Figures 4–7. In these, the pose docking ranks are plotted against the smina affinity scores. The Top-1 docking smina affinity scores are presented in Supplementary Table 2, and the average of the Top-5 docking smina affinity scores are presented in Supplementary Table 3.

Trends of docking rank & binding affinity

The plots of DiffDock protein-PFAs docking ranks versus smina affinity scores (Figures 4–7) show a general trend of the smina affinity scores becoming more negative as the docking improves with the top-rank docks having the top confidence scores. Because the smina affinity score is an approximation of the Gibbs free energy of the system [43], the more negative the smina affinity score becomes, the more stable the protein–PFAs complex being formed. Note that the smina affinity score was calculated after the docking inference steps, which implies that the trends exhibited in the rank-versus-affinity score plots (Figures 4–7) confirm how the diffusion generative model of DiffDock guides the inference of ligand pose and docking to a state consistent with the principles of molecular thermodynamics where a chemical system attains stability by seeking very negative Gibbs free energy [44].

The smina affinity score (kcal/mol) with the most negative values are close to the range of published experimental and simulation-based binding free energy of stable protein–ligand complex at around -8 kcal/mol [45]. The apparent randomness in the smina affinity scores in the sequence of the docking ranks is due to the probabilistic approach of the diffusion generative model in DiffDock sampling the ligand pose [13]. The convergence to optimal docking position results in the general trend of favoring a more negative smina score as the docking rank improves.

Comparison with related PFAs empirical results

It can be observed that many of the PFAs have very negative smina scores towards albumin and CBG, which indicates strong binding affinity (Supplementary Tables 2 & 3). On the other hand, almost all of the PFAs have weak binding affinity towards hemoglobin and alpha-1-antitrypsin (Supplementary Tables 2 & 3). Based on top-1 docking, PFHxS among all the PFAs has the strongest binding towards albumin, with smina score of -4.88 kcal/mol, and PFOS among all the PFAs has the strongest binding towards CBG with smina score of -6.88 kcal/mol. Also evident in Supplementary Table 3 is the varying standard deviation of the top-5 docking. Low standard deviation in smina scores of the top docking (top-5) indicates good convergence of the predictions, and high standard deviation in smina scores indicates bad convergence of the predictions.

Although studies have shown that glycated hemoglobin (HbA1c) has some positive correlation with PFOA concentration in blood [46,47], there has not been definitive reports of complex formation between hemoglobin and PFAs. The trends of PFAs–hemoglobin

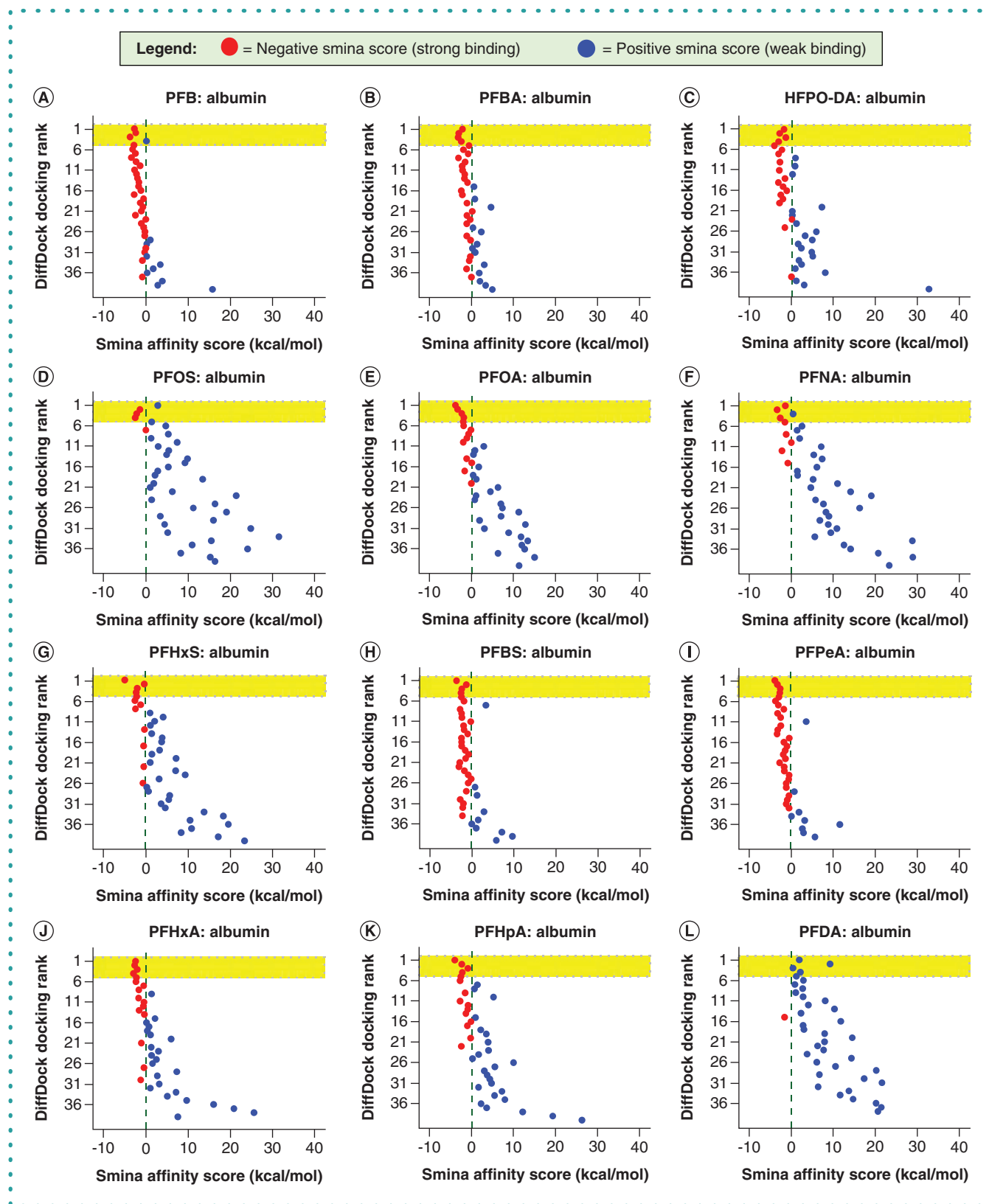


Figure 4. Molecular docking pose rank computed using DiffDock and the associated smina affinity score between albumin protein molecule and PFAs. Top-5 docking ranks are highlighted in yellow. (A) PFB, (B) PFBA, (C) HFPO-DA, (D) PFOS, (E) PFOA, (F) PFNA, (G) PFHxS, (H) PFBS, (I) PFPeA, (J) PFHxA, (K) PFHpA, (L) PFDA.

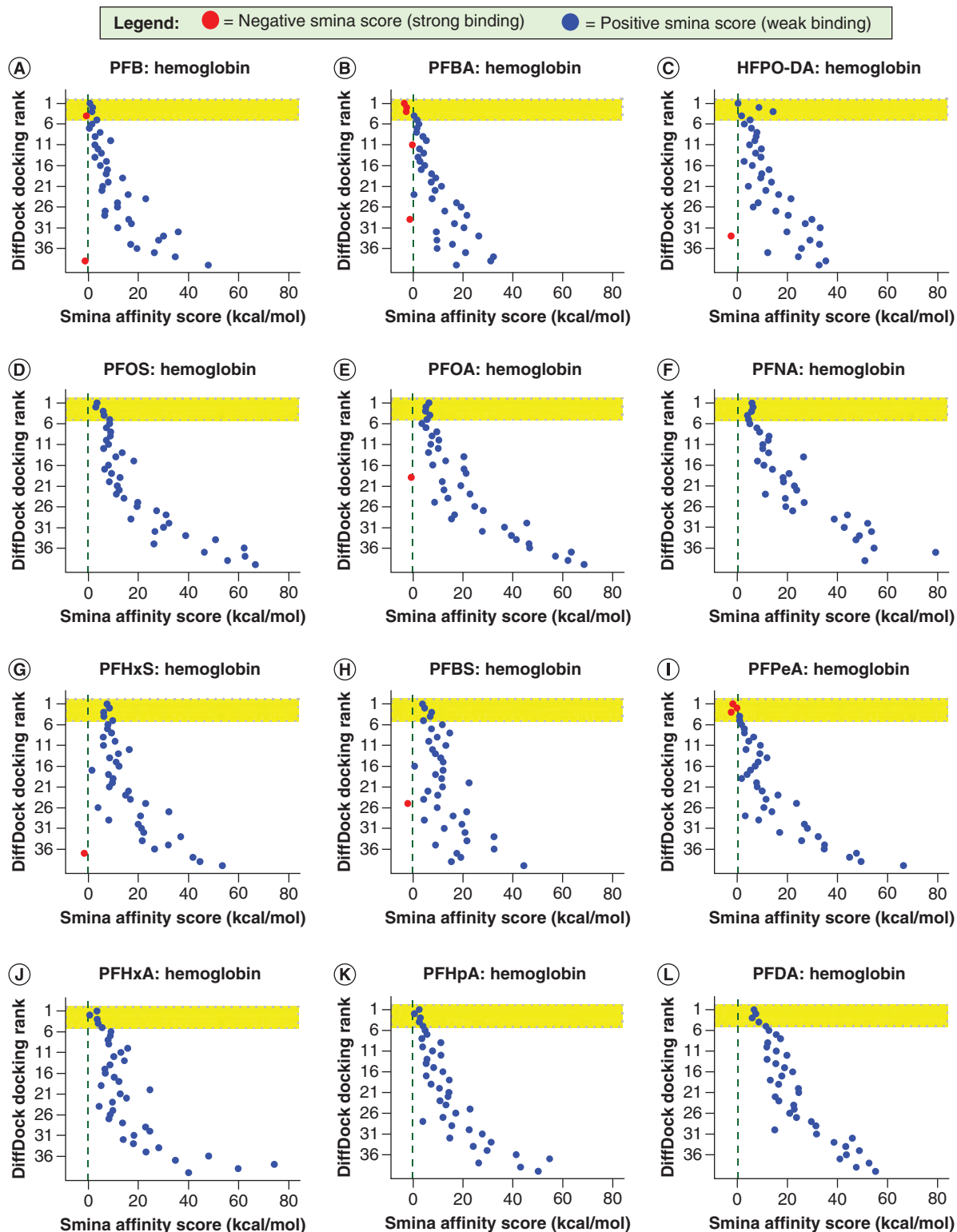


Figure 5. Molecular docking pose rank computed using DiffDock and the associated smina affinity score between hemoglobin protein molecule and PFAs. Top-5 docking ranks are highlighted in yellow. (A) PFB, (B) PFBA, (C) HFPO-DA, (D) PFOS, (E) PFOA, (F) PFNA, (G) PFHxS, (H) PFBS, (I) PFPeA, (J) PFHxA, (K) PFHpA, (L) PFDA.

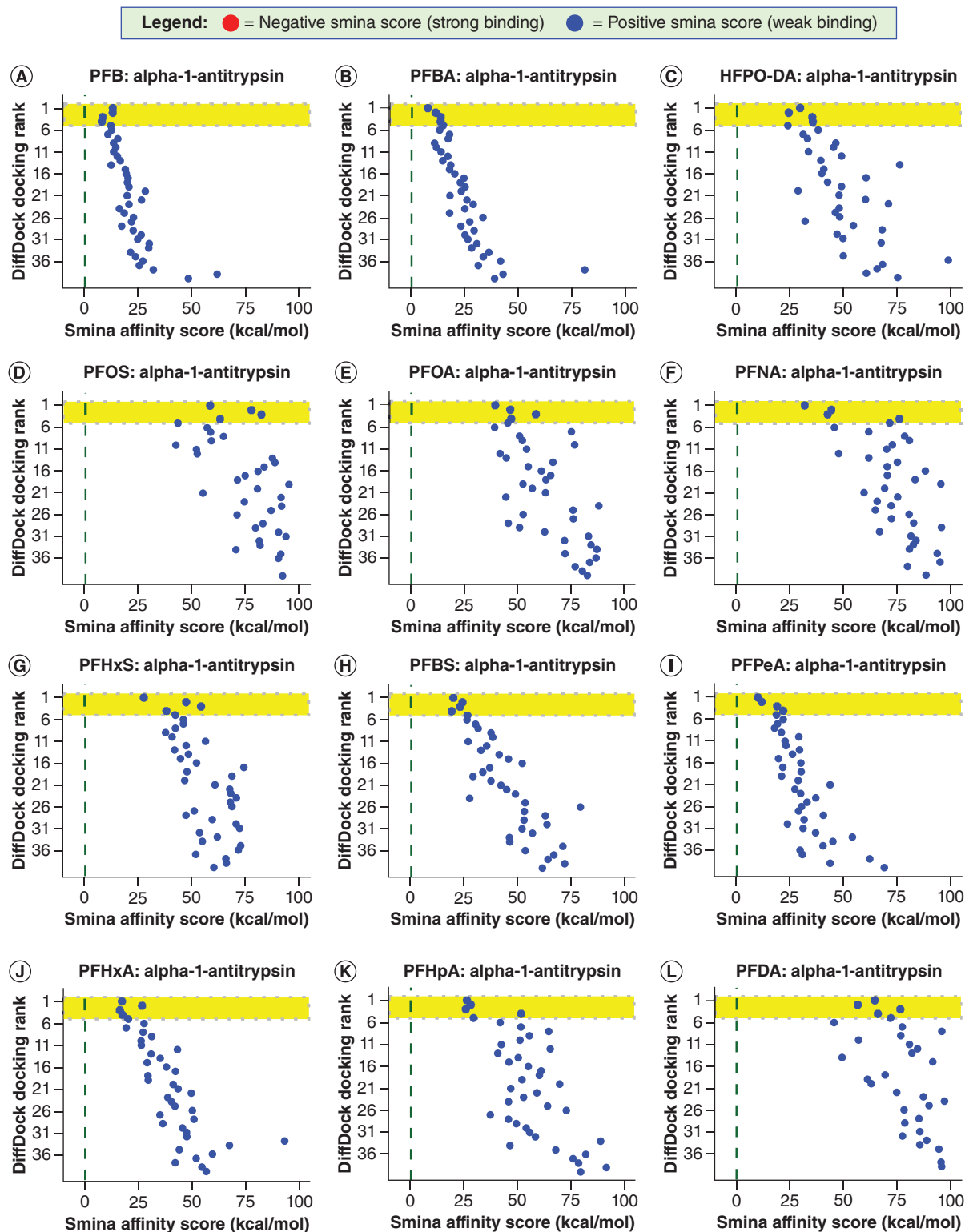


Figure 6. Molecular docking pose rank computed using DiffDock and the associated smina affinity score between alpha-1-antitrypsin protein molecule and PFAs. Top-5 docking ranks are highlighted in yellow. (A) PFB, (B) PFBA, (C) HFPO-DA, (D) PFOS, (E) PFOA, (F) PFNA, (G) PFHxS, (H) PFBS, (I) PFPeA, (J) PFHxA, (K) PFHpA, (L) PFDA.

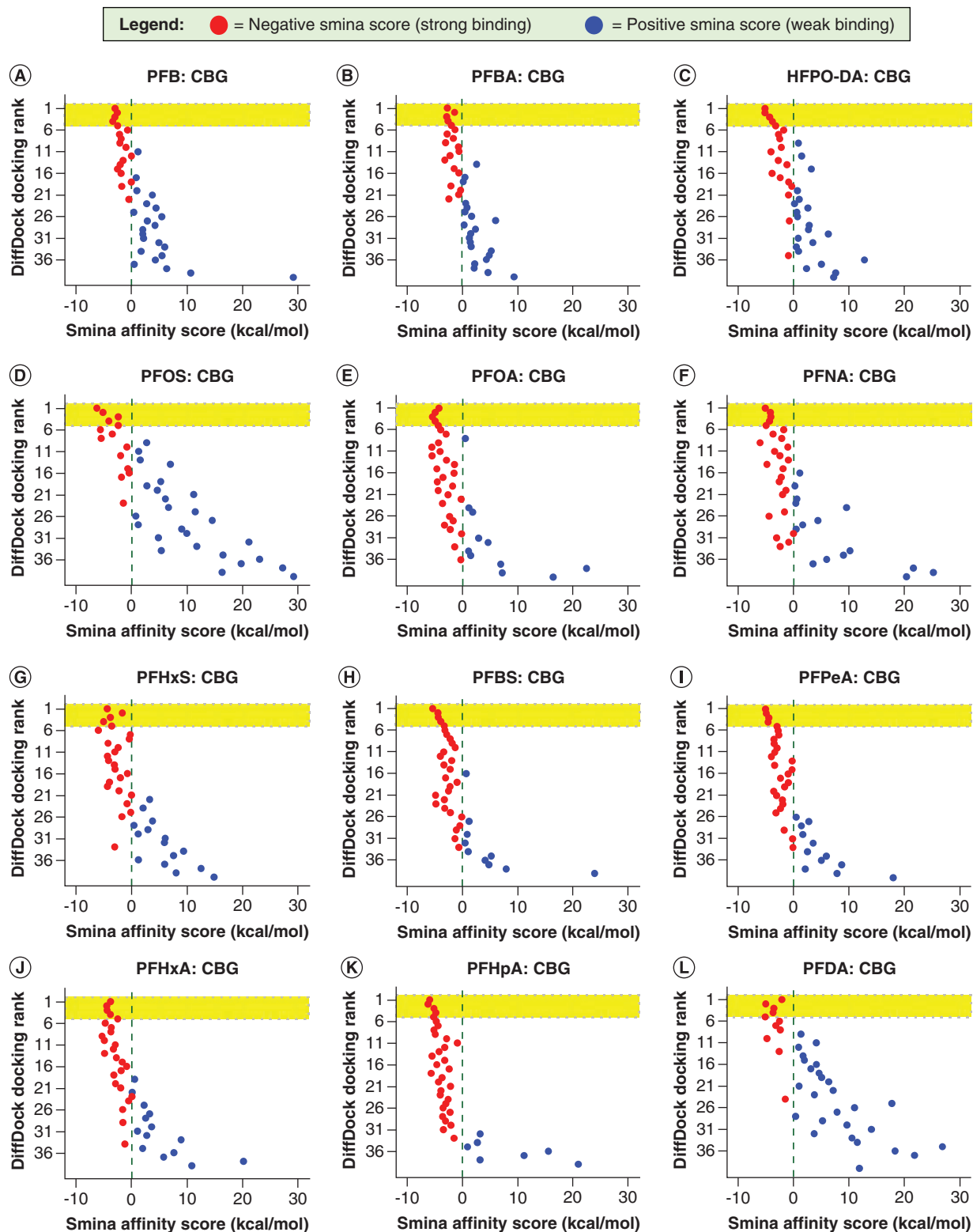


Figure 7. Molecular docking pose rank computed using DiffDock and the associated smina affinity score between corticosteroid-binding globulin protein molecule and PFAs. Top-5 docking ranks are highlighted in yellow. (A) PFB, (B) PFBA, (C) HFPO-DA, (D) PFOS, (E) PFOA, (F) PFNA, (G) PFHxS, (H) PFBS, (I) PFPeA, (J) PFHxA, (K) PFHpA, (L) PFDA.

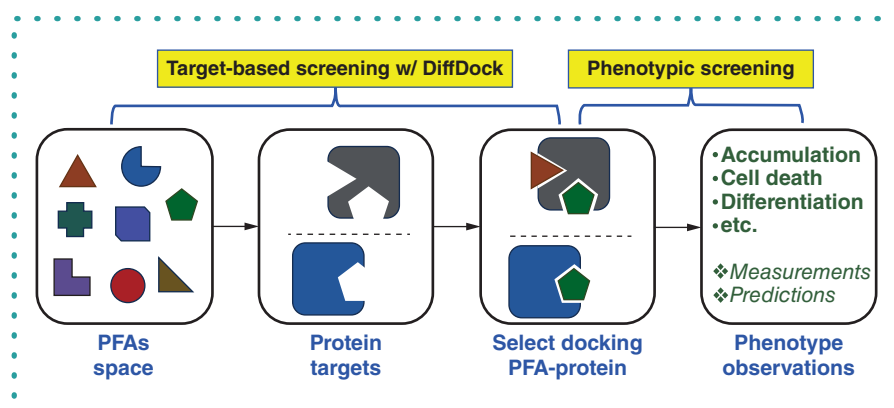


Figure 8. Integration of target-based screening and phenotypic screening of perfluoroalkyl and polyfluoroalkyl substances for their binding and effects on human proteins. The large pool of perfluoroalkyl and polyfluoroalkyl substances (PFAs) identified to have been released to the environment is first screened computationally using a generative AI algorithm such as DiffDock to narrow down the set of PFA–proteins of concern. Then the select PFA–protein pairs would be used to make measurements or predictions of phenotype properties.

complex docking shown in Figure 5 implies very weak binding affinity between these protein–ligand systems. This suggests that the determined correlation between PFAs and HbA1c may be the result of indirect mechanistic association.

The selection of alpha-1-antitrypsin for this study is because of the lack of studies on its chemistry with PFAs amid the fact that it has been recognized as a glycoprotein that may have multifunctional roles with therapeutic effects in many inflammatory and autoimmune diseases [48]. With this value, the potential of alpha-1-antitrypsin as a carrier of PFAs to other individuals via blood transfusion must be evaluated. The results in Figure 6 show that all PFAs tested have very weak binding affinity towards alpha-1-antitrypsin protein, as indicated by the positive smina scores. This may mean that the risk of carrying PFAs from one person to another with alpha-1-antitrypsin from blood is very low.

Significance of the current work

Reviews on the literature of PFAs have found that more than 4700 of these substances have been introduced into many of the products and ecosystems that are used in everyday life [49–51]. This large pool of PFAs poses an inherent challenge in conducting studies on their pathways in the environment and in the human body [49]. However, the closely related field of drug discovery has recently been making significant progress in solving the same problem of studying numerous possible compounds that may affect the human body [13]. One of the common steps in both study areas is the analysis of molecular docking of protein–ligand complexes. Hence, this work demonstrates the potential to accelerate analysis of PFA affinity to the human body using the generative AI algorithm DiffDock, which originated as a computation tool in drug discovery [13].

On top of screening the strong and weak binding of protein–PFA complexes, the use of a computational algorithm such as DiffDock may also help elucidate chemical pathways. An example case is that of PFOS, which has been reported by empirical measurement to strongly complex with albumin [8]. The DiffDock computation results do not suggest this finding (Figure 4D), and show that other proteins may be responsible for the strong binding of PFOS to blood components such as CBG, as shown in Figure 7D & Supplementary Table 2. With this intriguing result, the relevant literature of PFOS measurements in human blood was then reviewed and it was found that these measurements involve the physical separation of blood components before PFOS and other PFAs are measured [8]. The fractionation of human blood via ultracentrifuge used by Forsthuber *et al.* [8] can reach only 85–90% purity for albumin fraction [52]. This leads to the possibility that other proteins such as CBG are not completely separated from albumin and are the proteins carrying PFOS via complex formation (Figure 7D) and not albumin (Figure 4D). Another method used in blood fractionation is electrophoresis, but this method is not selective towards albumin, which migrates toward the anode but at a slower rate compared with the prealbumin proteins transthyretin (sometimes called thyroxine-binding prealbumin) and retinol-binding protein [53]. Hence, the current methods of separating albumin from the rest of the blood components may affect the measurements on whether PFAs are complexed with albumin. This analysis of PFOS trends illustrates the type of investigation that follows after molecular docking analysis is performed and can lead to more precise measurements of PFA complexes and their pathways in the human body.

The set of speculative statements mentioned in the previous paragraph is the type of inquiry we suggest in positioning the DiffDock approach as a step in the larger context of studying the effect of PFAs on human health. A proposed concept of integrating DiffDock into the study of PFAs in relation to human health is shown in Figure 8. DiffDock becomes a target-based screening tool in the initial stage of determining key pairs of PFA–protein systems that will be eventually used for phenotype observation studies. Such an *in silico* step with DiffDock follows similar implementation as in the area of drug discovery [54,55]. This approach reduces the pool of PFAs of

concern in relation to target proteins; thus can be used to screen PFAs from the known pool of PFAs released to the environment, which would then allow a comprehensive screening of PFAs of concern.

The predictions of *in silico* molecular docking must be ultimately verified empirically such as the computation of root-mean-square deviation between the predicted docking and the measured crystalline structure of protein–ligand docking [56]. Using *in silico* analysis such as the use of DiffDock solves a critical bottleneck, which is the task of narrowing down the set of candidate protein–ligand docking to feasible size amenable to empirical measurements [57]. The DiffDock algorithm's inference computations, which are the generative AI aspect of the algorithm, makes the prediction of the next best pocket to dock a ligand pose after learning from previous docking steps [13]. In essence, the DiffDock algorithm can initially narrow down the pool of PFA–protein complexes that can be experimentally studied and validated (Figure 8). The scope of the paper is limited to showing the potential of computational inferencing of docking using DiffDock.

Conclusion

Our study findings demonstrate the possibility to accelerate human blood protein–PFA molecular docking computations using the generative AI algorithm DiffDock.

Future perspective

With the demonstration of using the generative AI algorithm DiffDock to accelerate molecular docking computations of PFAs towards human blood proteins, we anticipate the integration of the algorithm as an initial stage target-based screening tool in the workflow studying the effects of PFAs on human health (Figure 8). This initial stage would narrow down the pool of PFAs of concern in relation to target proteins allowing for more efficient allocation of laboratory experimental time and resources. This concept may also be used in the process of designing new chemical substitutes to PFAs in which the DiffDock algorithm may predict binding affinities of such new chemical substitutes towards human blood proteins.

Executive summary

Background

- PFA accumulation in human blood has been an increasing concern due to their potential serious negative effects on human health.
- This work aims to demonstrate the capability of generative AI algorithm DiffDock, which was originally developed for drug discovery as a target-based screening tool for PFA binding to human blood proteins.

Materials & methods

- Molecular structures of four human blood proteins were downloaded from Protein Data Bank: albumin, hemoglobin, alpha-1-antitrypsin and corticosteroid-binding globulin (CBG).
- The SMILES file formats of 12 PFAs were collected from PubChem: PFB, PFBA, HFPO-DA, PFOS, PFOA, PFNA, PFHxS, PFBS, PFPeA, PFHxA, PFHpA and PFDA.
- The Python code implementation of the DiffDock algorithm was run to perform molecular docking and compute the docking scores of the PFAS towards the blood proteins.

Results & discussion

- The various PFAs exhibit varied levels of binding affinity towards human blood proteins.
- PFOS exhibits weak binding with albumin and strong binding affinity towards CBG.
- Of the four human blood proteins, albumin and CBG results in the most negative ΔG affinity scores, indicating strong protein–PFA binding while PFA complexes with hemoglobin and alpha-1-antitrypsin are weak.

Conclusions

- Generative AI algorithm DiffDock accelerates protein–PFA molecular docking computations.

Supplementary data

To view the supplementary data that accompany this paper please visit the journal website at: www.future-science.com/doi/suppl/10.2144/btn-2023-0070

Author contributions

Conceptualization, DLB Fortela, AP Mikolajczyk and W Sharp; methodology, DLB Fortela, AP Mikolajczyk and MR Carnes; software, DLB Fortela and MR Carnes; formal analysis, DLB Fortela, AP Mikolajczyk, W Sharp, E Revellame, R Hernandez, WE Holmes and MZ; data curation, DLB Fortela; project administration, DLB Fortela; funding acquisition, DLB Fortela, MR Carnes and M Zappi. All authors have read and agreed to the published version of the manuscript.

Acknowledgments

The authors thank the supportive staff and students of the Department of Chemical Engineering and the Energy Institute of Louisiana at the University of Louisiana at Lafayette. We are also grateful to LaSPACE for supporting students who pursue research in STEM projects.

Financial disclosure

This work was partially funded through the LURA program by the Louisiana Space Grant Consortium (LaSPACE) with sub-award number PO-0000168186 under the main NASA grant number 80NSSC20M0110. The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

Competing interests disclosure

The authors have no competing interests or relevant affiliations with any organization or entity with the subject matter or materials discussed in the manuscript. This includes employment, consultancies, honoraria, stock ownership or options, expert testimony, grants or patents received or pending, or royalties.

Writing disclosure

No writing assistance was utilized in the production of this manuscript.

Data availability statement

The Python codes used in implementing the DiffDock algorithm were based on the codes originally developed by Corso G, Stärk H, Jing B, Barzilay R, Jaakkola T [13]. The codes are made available as Jupyter Notebook files in the online repository of the paper via GitHub: https://github.com/dhanfort/Blood.Protein.PFAS_Docking.git [21]. The following supporting information are also available in the GitHub repository of the paper: Video S1: DiffDock inference on PFBA with albumin, and Video S1: DiffDock inference on PFOS with alpha-1-antitrypsin.

Open access

This work is licensed under the Attribution-NonCommercial-NoDerivatives 4.0 Unported License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>

References

Papers of special note have been highlighted as: ●● of considerable interest

1. EPA. PFAS strategic roadmap: EPA's commitments to action 2021–2024. <https://www.epa.gov/pfas/pfas-strategic-roadmap-epas-commitments-action-2021-2024>
2. Langenbach B, Wilson M. Per- and polyfluoroalkyl substances (PFAS): significance and considerations within the regulatory framework of the USA. *Int. J. Environ. Res. Public Health* 18(21), 1–17 (2021).
- 3. Andersen ME, Hagenbuch B, Apte U *et al.* Why is elevation of serum cholesterol associated with exposure to perfluoroalkyl substances (PFAS) in humans? A workshop report on potential mechanisms. *Toxicology* 459, 152845 (2021).
4. Patlewicz G, Richard Ann M, Williams Antony J *et al.* A chemical category-based prioritization approach for selecting 75 per- and polyfluoroalkyl substances (PFAS) for tiered toxicity and toxicokinetic testing. *Environ. Health Perspect.* 127(1), 014501 (2019).
5. Verner MA, Ngueta G, Jensen ET *et al.* A simple pharmacokinetic model of prenatal and postnatal exposure to perfluoroalkyl substances (PFASs). *Environ. Sci. Technol.* 50(2), 978–986 (2016).
6. NIH. Perfluoroalkyl and Polyfluoroalkyl Substances (PFAS). <https://newsinhealth.nih.gov/2022/03/perfluoroalkyl-polyfluoroalkyl-substances-pfas>
7. CDC. Per- and Polyfluoroalkyl Substances (PFAS) and Your Health: PFAS Blood Testing. <https://www.atsdr.cdc.gov/pfas/health-effects/blood-testing.html>
8. Forsthuber M, Kaiser AM, Granitzer S *et al.* Albumin is the major carrier protein for PFOS, PFOA, PFHxS, PFNA and PFDA in human plasma. *Environ. Int.* 137, 105324 (2020).
- 9. CDC. Per- and Polyfluoroalkyl Substances (PFAS) and Your Health: Estimating Levels of PFAS in Your Blood. <https://www.atsdr.cdc.gov/pfas/resources/estimating-pfas-blood.html#:~:text=The%20Centers%20for%20Disease%20Control%20and%20Prevention%20%28CDC%29,%28PFHxS%29%2C%20and%20perfluorononanoic%20acid%20%28PFNA%29%20in%20the%20blood>
10. Qiu Y, Myers DR, Lam WA. The biophysics and mechanics of blood from a materials perspective. *Nat. Rev. Mater.* 4(5), 294–311 (2019).
11. Castillo DJ, Rifkin RF, Cowan DA, Potgieter M. The healthy human blood microbiome: fact or fiction? *Front. Cell. Infect. Microbiol.* 9(148), 9 (2019).
12. Sohn E. Diagnosis: frontiers in blood testing. *Nature* 549(7673), S16–S18 (2017).
13. Corso G, Stärk H, Jing B, Barzilay R, Jaakkola T. DiffDock: diffusion steps, twists, and turns for molecular docking. International Conference on Learning Representations (ICLR), Kigali, Rwanda, <https://iclr.cc/virtual/2023/poster/11750> (2023).
- 14. Creators of the DiffDock algorithm demonstrate how DiffDock performs with benchmark datasets, and publishes the Python codes used to run the algorithm.
15. Aithani L, Alcaide E, Bartunov S *et al.* Advancing structural biology through breakthroughs in AI. *Curr. Opin. Struct. Biol.* 80, 102601 (2023).
16. Bertram HM, Westbrook J, Feng Z *et al.* The Protein Data Bank. *Nucleic Acids Res.* 28(1), 235–242 (2000).
17. Song Y, Sohl-Dickstein J, Kingma DP, Kumar A, Ermon S, Poole B. Score-based generative modeling through stochastic differential equations. *arXiv* doi:<https://doi.org/10.48550/arXiv.2011.13456> (2021).
- 18. Develops and demonstrates the capability stochastic differential equations in improving score-based generative AI predictions eventually used by others for the DiffDock algorithm.
19. Van Rossum G. Python tutorial: technical report CS-R9526 (1995). <https://ir.cwi.nl/pub/5007/05007D.pdf>
20. Foundation PS. Python (2023). <https://www.python.org>
21. Google. Welcome to Colaboratory! (2018). <https://colab.research.google.com>
22. Kluyver T, Ragan-Kelley B, Perez F *et al.* Jupyter Notebooks – a publishing format for reproducible computational workflows. In: *Positioning and Power in Academic Publishing: Players, Agents and Agendas*. Schmidt FLA (Ed.). IOS Press, 87–90 (2016).
23. Paper title. <https://anonymous.4open.science/>
24. Pettersen EF, Goddard TD, Huang CC *et al.* UCSF ChimeraX: structure visualization for researchers, educators, and developers. *Protein Sci.* 30(1), 70–82 (2021).
25. Goddard TD, Huang CC, Meng EC *et al.* UCSF ChimeraX: meeting modern challenges in visualization and analysis. *Protein Sci.* 27(1), 14–25 (2018).
26. O'boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR. Open Babel: an open chemical toolbox. *J. Cheminformatics* 3(1), 33 (2011).
27. Smith GS, Walter GL, Walker RM. Chapter 18 – Clinical pathology in non-clinical toxicology testing. In: *Haschek and Rousseaux's Handbook of Toxicologic Pathology (Third Edition)*. Haschek WM, Rousseaux CG, Wallig MA (Eds.). Academic Press, MA, USA, 565–594 (2013).

26. Perutz MF, Rossmann MG, Cullis AF, Muirhead H, Will G, North ACT. Structure of haemoglobin: a three-dimensional fourier synthesis at 5.5-Å. Resolution, obtained by x-ray analysis. *Nature* 185(4711), 416–422 (1960).
27. Gettins PGW. Serpin structure, mechanism, and function. *Chem. Rev.* 102(12), 4751–4804 (2002).
28. Jirikowski GF, Rodewald A, Sivukhina E, Caldwell J. Corticosteroid binding globulin. In: *Reference Module in Neuroscience and Biobehavioral Psychology*. Elsevier (2017).
29. Sugio S, Mochizuki S, Noda M, Kashima A. Crystal structure of human serum albumin. *Protein Data Bank* doi:https://doi.org/10.2210/pdb1ao6/pdb (1998).
30. Sugio S, Kashima A, Mochizuki S, Noda M, Kobayashi K. Crystal structure of human serum albumin at 2.5 Å resolution. *Protein Eng. Des. Selec.* 12(6), 439–446 (1999).
31. Tame J, Vallone B. Deoxy human hemoglobin. *Protein Data Bank* doi:https://doi.org/10.2210/pdb1a3n/pdb (1998).
32. Tame JRH, Vallone B. The structures of deoxy human haemoglobin and the mutant Hb Tyr[alpha]42His at 120 K. *Acta Crystallogr. D* 56(7), 805–811 (2000).
33. Elliott PR, Pei XY, Dafforn T, Read RJ, Carrell RW, Lomas DA. 2.0 Angstrom structure of intact alpha-1-antitrypsin: a canonical template for active serpins. *Protein Data Bank* doi:https://doi.org/10.2210/pdb1qlp/pdb (1999).
34. Elliott PR, Pei XY, Dafforn TR, Lomas DA. Topography of a 2.0 Å structure of α1-antitrypsin reveals targets for rational drug design to prevent conformational disease. *Protein Sci.* 9(7), 1274–1281 (2000).
35. Zhou A, Wei Z, Read RJ. Crystal structure of the reactive loop cleaved corticosteroid binding globulin complexed with cortisol. *Protein Data Bank* doi:https://doi.org/10.2210/pdb2vdy/pdb (2007).
36. Zhou A, Wei Z, Stanley PLD, Read RJ, Stein PE, Carrell RW. The S-to-R transition of corticosteroid-binding globulin and the mechanism of hormone release. *J. Mol. Biol.* 380(1), 244–251 (2008).
37. Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inform. Comp. Sci.* 28(1), 31–36 (1988).
38. Kim S, Chen J, Cheng T et al. PubChem 2023 update. *Nucleic Acids Research* 51(D1), D1373–D1380 (2023).
39. Longpré D, Lorusso L, Levicki C, Carrier R, Cureton P. PFOS, PFOA, LC-PFCAS, and certain other PFAS: a focus on Canadian guidelines and guidance for contaminated sites management. *Enviro. Technol. Innov.* 18, 100752 (2020).
40. Schulz K, Silva MR, Klaper R. Distribution and effects of branched versus linear isomers of PFOA, PFOS, and PFHxS: a review of recent literature. *Sci. Total Environ.* 733, 139186 (2020).
41. Koes DR, Baumgartner MP, Camacho CJ. Lessons learned in empirical scoring with smina from the CSAR 2011 benchmarking exercise. *J. Chem. Inf. Model* 53(8), 1893–1904 (2013).
- Demonstrates the use of smina scoring function for molecular docking assessment of ligands and proteins.
42. Trott O, Olson AJ. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.* 31(2), 455–461 (2010).
43. Hill AD, Reilly PJ. Scoring functions for AutoDock. *Methods Mol. Biol.* 1273, 467–474 (2015).
44. Smith JM, Van Ness H, Abbott M, Swihart M. *Introduction to Chemical Engineering Thermodynamics*. McGraw-Hill, NY, USA, 9 (2022).
45. Woo HJ, Roux B. Calculation of absolute protein–ligand binding free energy from computer simulations. *Proc. Natl Acad. Sci. USA* 102(19), 6825–6830 (2005).
46. Cakmak S, Lukina A, Karthikeyan S, Atlas E, Dales R. The association between blood PFAS concentrations and clinical biochemical measures of organ function and metabolism in participants of the Canadian Health Measures Survey (CHMS). *Sci. Total Environ.* 827, 153900 (2022).
47. Zare Jedd M, Dalla Zuanna T, Barbieri G et al. Associations of perfluoroalkyl substances with prevalence of metabolic syndrome in highly exposed young adult community residents – a cross-sectional study in Veneto Region, Italy. *Int. J. Environ. Res. Public Health* 18(3), (2021).
48. Kim M, Cai Q, Oh Y. Therapeutic potential of alpha-1 antitrypsin in human disease. *Ann. Pediatr. Endocrinol. Metab.* 23(3), 131–135 (2018).
49. Panieri E, Baralic K, Djukic-Cosic D, Buha Djordjevic A, Saso L. PFAS molecules: a major concern for the human health and the environment. *Toxics* 10(2), (2022).
50. Steenland K, Winquist A. PFAS and cancer, a scoping review of the epidemiologic evidence. *Environ. Res.* 194, 110690 (2021).
51. Fenton SE, Ducatman A, Boobis A et al. Per- and polyfluoroalkyl substance toxicity and human health review: current state of knowledge and strategies for informing future research. *Environ. Toxicol. Chem.* 40(3), 606–630 (2021).
52. Harris JR. *Blood separation and plasma fractionation*. Wiley-Liss, NY, USA, 497 (1991).
53. Bertholf RL. Proteins and albumin. *Lab. Med.* 45(1), e25–e41 (2014).
54. Iwata H, Kojima R, Okuno Y. An *in silico* approach for integrating phenotypic and target-based approaches in drug discovery. *Mol. Inform.* 39(1–2), e1900096 (2020).
55. Lauria A, Bonsignore R, Bartolotta R, Perricone U, Martorana A, Gentile C. Drugs polypharmacology by *in silico* methods: new opportunities in drug discovery. *Curr. Pharm. Des.* 22(21), 3073–3081 (2016).
56. Bell EW, Zhang Y. DockRMSD: an open-source tool for atom mapping and RMSD calculation of symmetric molecules through graph isomorphism. *J. Cheminform.* 11(1), 40 (2019).
57. Stokes JM, Yang K, Swanson K et al. A deep learning approach to antibiotic discovery. *Cell* 180(4), 688–702.e613 (2020).