Development and Application of *In Silico* Machine Learning Models for the Investigation of Halogen Bonding in the Drug Discovery Process

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät

der Eberhard Karls Universität Tübingen

zur Erlangung des Grades eines

Doktors der Naturwissenschaften

(Dr. rer. nat.)

vorgelegt von

Marc Uwe Engelhardt

aus Ludwigsburg

Tübingen

2026

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:	17.04.2026
Dekan:	Prof. Dr. Thilo Stehle
1. Berichterstatter/-in:	Prof. Dr. Frank M. Böckler
2. Berichterstatter/-in:	PD Dr. Thomas E. Exner

Selbstständigkeitserklärung

Ich, Marc Uwe Engelhardt, erkläre hiermit, dass ich die zur Promotion eingereichte Arbeit mit dem Titel: „*Development and Application of In Silico Machine Learning Models for the Investigation of Halogen Bonding in the Drug Discovery Process“* selbstständig verfasst, nur die angegebenen Quellen und Hilfsmittel benutzt und wörtlich oder inhaltlich übernommene Stellen als solche gekennzeichnet habe. Ich versichere an Eides statt, dass diese Angaben wahr sind und dass ich nichts verschwiegen habe. Eine detaillierte Abgrenzung meiner eigenen Leistungen von den Beiträgen meiner Kooperationspartnern habe ich vorgenommen.

Declaration of Authorship

I, Marc Uwe Engelhardt, hereby declare that I have produced the work entitled „*Development and Application of In Silico Machine Learning Models for the Investigation of Halogen Bonding in the Drug Discovery Process“*, submitted for the award of a doctorate, on my own, have used only the sources and aids indicated and have marked passages included from other works, whether verbatim or in content, as such. I swear upon oath that these statements are true and that I have not concealed anything. I have clearly and in detail distinguished my own contributions from those of my cooperation partners.

Tübingen, den 15. Januar 2026 Marc Uwe Engelhardt

Acknowledgements

Whatever I have accomplished in this dissertation, I could not have done it alone. Many individuals have supported, guided, and inspired me throughout this journey, and I am deeply grateful for all contributions.

First and foremost, I am deeply grateful to my supervisor, **Prof. Dr. Frank M. Böckler**. Your constant help and support throughout my time as a PhD student were invaluable. You always took the time to discuss results and challenges and consistently provided insightful and constructive feedback on every topic. Thank you for giving me the opportunity to work in your group in a highly engaging research field, and for everything I learned during my time there, not only about research, but also about navigating the world of academia and just about several things under the sun. Thanks to you, I also had the opportunity to attend conferences across Europe, broadening both my academic and personal horizons.

I would like to thank **Dr. Thomas E. Exner** for being the second examiner of my thesis. Thank you for your time and expertise surveying my thesis.

Many thanks go to **Markus Zimmermann**, who accompanied me throughout my entire PhD and taught me countless aspects of computational chemistry, drug design in general, and halogen bonding in particular. You were always available for discussions and advice on virtually any topic. Over the years, you became not only an invaluable colleague but also a very good friend, and I am grateful that I can reach out to you at almost any time. I also thank you for the occasional emergency Tuplas.

Finn Mier joined our group in 2023 and has been my direct desk neighbor ever since (until the final act of treachery). I am very grateful for your constant support, particularly with programming and machine learning. Together with Markus, you were the best “rubber ducks” I could have asked for. Over time, you became a good friend both in and outside of work. Thank you for your never-ending supply of dad jokes and for making every day more enjoyable.

I would like to thank **Martin Schwer** and **Jason Stahlecker**, who became very good friends both in and outside of work. Our daily after-lunch darts games always provided a relaxing and recovering break. Moreover, it was, and still is, great that the three of us joined a local darts club together, taking part in tournaments and matches and continuously supporting one another.

I want to thank all members of the **Böckler Lab** in general, past and present, and especially **Janosch, Larissa, and Theresa.** I could not think of a better group I could have been in.

Beyond academia, I would like to thank my best mate, **Niklas**. Even though we no longer live close to each other, I know I can always count on you, no matter what. Your constant support and friendship mean more to me than I can express.

Thanks to all my friends and former housemates of the trap house, particularly **Benni, Jaques, Matze, and Dome**. Thank you for always supporting me and keeping me grounded through this process and for several days and evenings to relief stress.

A heartful thanks goes to my friends from Ludwigsburg, **Amti, Fezi, Alex, and Ruben**. Thank you for your continued support and our annual guys' trips.

I would like to thank my Darts-Team, **TSV Game Over Oferdingen**, and all members and players of it. Training, as well as parties, tournaments and league matches are always insanely nice and fun with you.

I would also like to thank my future parents-in-law, **Ute** and **Klaus**, as well as my future sister-in-law, **Maren**, and **Ari**.

Beyond that, I am deeply grateful to my mom **Monika**, my dad **Uwe** and his wife **Katrin** for their love and support. Your presence has given me the security to pursue my own path.

Finally, I would like to thank the love of my life, my fiancée **Jana**. Without you, none of this would have been possible. You encouraged me whenever things became difficult. With your exceptional organizational skills, you essentially helped coordinate my work and gave me the necessary pushes when they were needed. You were my safe haven, the person I could always rely on, unconditionally, during a genuinely challenging time. You constantly give me strength and motivation to overcome every challenge.

„Ich weiß nicht, wie ich der Welt erscheine,
aber mir selbst komme ich vor wie ein Knabe,
der am Meeresufer spielt und sich erfreut,
einen glatteren Kieselstein oder eine schönere Muschel zu finden,
während der große Ozean der Wahrheit unerforscht vor mir liegt.“
Isaac Newton (1643 - 1727)

Contents

List of Abbreviations

2D	Two-dimensional
AI	Artifical intelligence
ANN	Artificial neural network
aug-cc-pVTZ	Correlation-consistent polarized valence triple-ζ basis set augmented with a diffuse function
aug-cc-pVQZ	Correlation-consistent polarized valence quadruple-ζ basis set augmented with a diffuse function
B3LYP	Becke, three-parameter, Lee-Yang-Parr functional
Br	Bromine
BSSE	Basis set superposition error
CBS	Complete basis set
CC	Coupled cluster
cc-pVTZ	Correlation-consistent polarized valence triple-ζ basis set
cc-pVQZ	Correlation-consistent polarized valence quadruple-ζ basis set
CCSD(T)	Coupled cluster with single, double, and perturbative triple excitations
CI	Configuration interaction
Cl	Chlorine
CPU hours	Central processing unit hours
D3	Dispersion correction type 3
DFT	Density functional theory
DYRK1A	Dual-specificity tyrosine phosphorylation-regulated kinase 1A
GGA	Generalized gradient approximation
HF	Hartree-Fock
I	Iodine
KRR	Kernel ridge regression
LeakyReLU	Leaky rectified linear unit
M06-2X	Minnesota 06 hybrid functional with double Hartree-Fock exchange
MD	Mahalanobis distance
MD	Molecular dynamics
ML	Machine learning
MM	Molecular mechanics

MPn	Møller-Plesset perturbation method of order n
MSE	Mean squared error
NN	Neural network
PDB	Protein Data Bank
Post-HF	Post-Hartree-Fock
PP	Pseudopotential
QM	Quantum mechanical
R	Rest group
R^2	Coefficient of determination
ReLU	Rectified linear unit
RF	Random forest
RMSD	Root mean squared deviation
RMSE	Root mean squared error
SCS-MP2	Spin-component-scaled MP2
SVM	Support vector machine
Tanh	Hyperbolic tangent
TPSS	Tao, Perdew, Staroverov, and Scuseria exchange functional
TZVP	Valence triple-ζ basis set with one set of polarization function
TZVPP	Valence triple-ζ basis set with two sets of polarization functions
TZVPPD	Valence triple-ζ basis set with two sets of polarization functions and diffuse function
X	Halogen
XB	Halogen bond
XWH	Halogen-water-hydrogen bridge

Abstract

This dissertation comprises a series of studies that collectively elucidate the nature and modeling of halogen-mediated, noncovalent interactions through a combination of quantum mechanical (QM) and data-driven approaches. In the first part, a QM investigation of halogen···water interactions is conducted, with emphasis on their energetic and structural characteristics in biologically relevant environments such as protein binding sites. The second, and more extensive part of the thesis focuses on halogen···π interactions, integrating high-level QM calculations with machine learning (ML) approaches.

The first study investigated halogen···water interactions. Starting from a distinct iodine···water contact in a solved protein crystal structure, a comprehensive QM and database analysis was conducted and revealed that these interactions, though moderate in strength, are structurally well-defined and follow systematic trends across the halogen series. Chlorine was found to form flexible, mixed halogen-hydrogen-bonding arrangements, while iodine engaged in highly directional σ-hole interactions with water oxygen lone pairs.

The second study provided a quantitative benchmarking of halogen···π interactions and established MP2/TZVPP as the most balanced QM method for their description. This level of theory achieved near-reference accuracy with a root-mean-square deviation of approximately 1 kJ/mol relative to CCSD(T)/CBS data, ensuring consistency for subsequent modeling studies.

The third study introduced neural network models trained on high-level QM data to predict halogen···π interaction energies. The models reproduced MP2-level energies with excellent agreement ($R^2 \approx 0.998$) and achieved an approximate eight-order-of-magnitude (10^8) reduction in computational cost. Validation against both random and protein-derived geometries confirmed robust generalization within the σ-hole interaction domain.

The fourth study extended this QM-AI framework to include halogen···π interactions with phenol, imidazole, and indole systems, representing the aromatic side chains of tyrosine, histidine, and tryptophan. The extended models maintained near-MP2 accuracy across all systems ($R^2 \approx 0.99$) and demonstrated successful transferability to protein-derived geometries, confirming their scalability to chemically diverse

environments. Furthermore, the model's scalability and adaptability to new chemical environments was demonstrated by incorporating additional data and retraining, which led to improved performance.

Together, these findings establish a coherent and transferable framework for accurate and efficient modeling of halogen-mediated noncovalent interactions across varied molecular contexts.

Zusammenfassung

Diese Dissertation umfasst eine Reihe von Studien, die gemeinsam das Verständnis halogenvermittelter, nichtkovalenter Wechselwirkungen vertieft und deren Modellierung durch eine Kombination aus quantenmechanischer und datengetriebener Ansätze voranbringt. Im ersten Teil wird eine quantenmechanische (QM) Untersuchung von Halogen···Wasser-Wechselwirkungen durchgeführt, wobei der Schwerpunkt auf deren energetischen und strukturellen Eigenschaften in biologisch relevanten Umgebungen wie Proteinbindungstaschen liegt. Der zweite, umfangreichere Teil konzentriert sich auf Halogen···π-Wechselwirkungen und kombiniert hochpräzise, quantenmechanische Berechnungen mit Methoden des maschinellen Lernens (ML).

In der ersten Studie wurden Halogen···Wasser-Wechselwirkungen untersucht. Ausgangspunkt war ein eindeutiger Iod···Wasser-Kontakt in einer von uns aufgeklärten Proteinkristallstruktur. Eine umfassende quantenmechanische und datenbankgestützte Analyse zeigte, dass diese Wechselwirkungen zwar energetisch moderat, jedoch strukturell klar definiert sind und systematische Trends entlang der Halogenreihe aufweisen. Chlor bildet bevorzugt flexible, jedoch gemischte Halogen-Wasserstoffbrücken-Anordnungen, während Iod stark gerichtete σ-Loch-Wechselwirkungen mit den Elektronenpaaren des Wassersauerstoffs eingeht.

Die zweite Studie lieferte einen quantitativen Maßstab für Halogen···π-Wechselwirkungen und identifizierte MP2/TZVPP als die ausgewogenste quantenmechanische Methode zu deren Beschreibung. Dieses Berechnungsniveau erreichte nahezu Referenzgenauigkeit mit einer mittleren quadratischen Abweichung von etwa 1 kJ/mol gegenüber CCSD(T)/CBS-Daten und gewährleistete somit eine verlässliche Grundlage für die nachfolgenden Modellierungsstudien.

In der dritten Studie wurden auf hochpräzisen, quantenmechanischen Daten trainierte neuronale Netzwerke entwickelt, um Halogen···π-Wechselwirkungsenergien vorherzusagen. Die Modelle reproduzierten die MP2-Energien mit ausgezeichneter Übereinstimmung ($R^2 \approx 0.998$) und ermöglichten eine Reduktion der Rechenzeit um etwa acht Größenordnungen (10^8). Validierungen anhand zufällig generierter sowie aus Proteinstrukturen abgeleiteter Geometrien bestätigten eine robuste Generalisierungsfähigkeit innerhalb des σ-Loch-Wechselwirkungsbereichs.

Die vierte Studie erweiterte das QM-AI-System auf Halogen···π-Wechselwirkungen mit Phenol-, Imidazol- und Indol-Systemen, die den aromatischen Seitenketten der Aminosäuren Tyrosin, Histidin und Tryptophan entsprechen. Die erweiterten Modelle erreichten über alle Systeme hinweg nahezu MP2-genaue Vorhersagen ($R^2 \approx 0.99$) und zeigten eine erfolgreiche Übertragbarkeit auf Protein-abgeleitete Geometrien, was ihre Skalierbarkeit auf chemisch vielfältige Umgebungen bestätigte. Darüber hinaus wurden die Skalierbarkeit und Anpassungsfähigkeit des Modells an neue chemische Bedingungen durch die Einbeziehung zusätzlicher Daten und erneutem Training demonstriert, was zu einer Leistungssteigerung führte.

Insgesamt etablieren diese Studien ein konsistentes und übertragbares Gerüst zur präzisen und effizienten Modellierung halogenvermittelter, nichtkovalenter Wechselwirkungen in unterschiedlichen molekularen Kontexten.

List of Publications and Contributions

1. **Halogen Bonding on Water ─ A Drop in the Ocean?**
   Marc U. Engelhardt, Markus O. Zimmermann, Marcel Dammann, Jason Stahlecker, Antti Poso, Thales Kronenberger, Conrad Kunick, Thilo Stehle, Frank M. Boeckler;
   Journal of Chemical Theory and Computation, 2024,
   DOI: 10.1021/acs.jctc.4c00834

   Contributions:
   The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript. I, together with F.M.B. conceptualized the overall study. F.M.B. and M.O.Z. envisioned the research. M.D. and J.S. prepared the protein, conducted protein crystallization experiments, and performed data reduction and structure refinement. C.K. reviewed the manuscript and provided the compound 2-cyclopentyl-7-iodo-1*H*-indole-3-carbonitrile. I performed all QM calculations, wrote applications for PDB analysis, conducted PDB analysis and prepared the corresponding visualizations. M.O.Z. contributed to developing the computational strategy. T.K. and A.P. helped refine the concept and provided comments on the manuscript. I prepared the original draft. I, together with M.O.Z. and F.M.B., reviewed, edited, and finalized the manuscript.

2. **Comparison of QM Methods for the Evaluation of Halogen-π Interactions for Large-Scale Data Generation**
   Marc U. Engelhardt, Markus O. Zimmermann, Finn Mier, Frank M. Boeckler;
   Journal of Chemical Theory and Computation, 2025,
   DOI: 10.1021/acs.jctc.5c00456

   Contributions:
   The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript. I, together with F.M.B., conceptualized the overall study and envisioned the research. I performed all QM calculations, gathered all results and prepared the corresponding visualizations. M.O.Z. and F.M. contributed to developing the computational strategy and provided comments on the manuscript. I prepared the original draft. I, M.O.Z., F.M., and F.M.B., reviewed, edited, and finalized the manuscript.

3. **A QM-AI Approach for the Acceleration of Accurate Assessments of Halogen-π Interactions by Training Neural Networks**
   Marc U. Engelhardt, Finn Mier, Markus O. Zimmermann, Frank M. Boeckler;
   Journal of Chemical Information and Modeling, 2025,
   DOI: 10.1021/acs.jcim.5c02136

   Contriubitions:
   The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript. I, together with F.M.B., envisioned the research. I performed all QM calculations, developed all scripts for machine learning and evaluation processes, gathered all results and prepared the corresponding visualizations. I also prepared the original draft. F.M. contributed to developing the computational strategy. F.M. and M.O.Z. provided comments on the manuscript. I, M.O.Z., F.M., and F.M.B. reviewed, edited, and finalized the manuscript.

4. **Expanding and Enhancing Neural Network-Based QM-AI Models for the Accurate Prediction of Halogen-π Interaction Energies in Protein Contexts**

   Marc U. Engelhardt, Finn Mier, Markus O. Zimmermann, Frank M. Boeckler; *Journal of Chemical Information and Modeling*, 2025,
   DOI: 10.1021/acs.jcim.5c03249

   Contriubitions:

   The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript. I, together with F.M.B., envisioned the research. I conducted the majority of the QM calculations, developed all machine learning and evaluation scripts, assembled the complete data set of results, and prepared all visualizations. M.O.Z. performed a small subset of the QM calculations. I also prepared the original draft. F.M. contributed to developing the computational strategy and provided comments on the manuscript. I, F.M., M.O.Z, and F.M.B. reviewed, edited, and finalized the manuscript.

Accepted articles related to this dissertation

Principles and Applications of CF2X Moieties as Unconventional Halogen Bond Donors in Medicinal Chemistry, Chemical Biology, and Drug Discovery

Sebastian Vaas, Markus O. Zimmermann, Dieter Schollmeyer, Jason Stahlecker, **Marc U. Engelhardt**, Janosch Rheinganz, Bernhard Drotleff, Matthias Olfert, Michael Lämmerhofer, Markus Kramer, Thilo Stehle, Frank M. Boeckler;

Journal of Medicinal Chemistry, 2023, DOI: 10.1021/acs.jmedchem.3c00634

Rights & Permissions for Partial and Full Use of Text and Images from Journal Articles

1. Halogen Bonding on Water – A Drop in the Ocean?

AMERICAN CHEMICAL SOCIETY LICENSE
TERMS AND CONDITIONS

Apr 20, 2026

This Agreement between Marc U. Engelhardt ("You") and American Chemical Society ("American Chemical Society") consists of your license details and the terms and conditions provided by American Chemical Society and Copyright Clearance Center.

License Number	6253100849519
License date	Apr 20, 2026
Licensed Content Publisher	American Chemical Society
Licensed Content Publication	Journal of Chemical Theory and Computation
Licensed Content Title	Halogen Bonding on Water—A Drop in the Ocean?
Licensed Content Author	Marc U. Engelhardt, Markus O. Zimmermann, Marcel Dammann, et al
Licensed Content Date	Dec 1, 2024
Licensed Content Volume	20
Licensed Content Issue	23
Volume number	20
Issue number	23
Type of Use	Thesis/Dissertation
Requestor type	Author (original work)
Format	Print and Electronic
Portion	Full article
Title of new work	Development and Application of In Silico Machine Learning Models for the Investigation of Halogen Bonding in the Drug Discovery Process
Institution name	Eberhard Karls Universität Tübingen
Expected presentation date	Apr 2026
The Requesting Person / Organization to Appear on the License	Marc U. Engelhardt
Requestor Location	Mr. Marc Engelhardt Auf der Morgenstelle 8 Tuebingen, 72076 Germany
Order reference number	ME_20_04_26
Payment Type	Invoice
Email Address	marc.engelhardt@uni-tuebingen.de
Billing Address	Mr. Marc Engelhardt Auf der Morgenstelle 8 Tuebingen, Germany 72076
Total	0.00 EUR
Terms and Conditions	

American Chemical Society's Policy on Thesis and Dissertations

If your university requires you to obtain permission, you must use the RightsLink permission system.
See RightsLink instructions at http://pubs.acs.org/page/copyright/permissions.html.

This is regarding request for permission to include **your** paper(s) or portions of text from **your** paper(s) in your thesis. Permission is now automatically granted; please pay special attention to the **implications** paragraph below. The Copyright Subcommittee of the Joint Board/Council Committees on Publications approved the following:

> Copyright permission for published and submitted material from thesis and dissertations
> ACS extends blanket permission to students to include in their thesis and dissertations their own articles, or portions thereof, that have been published in ACS journals or submitted to ACS journals for publication, provided that the ACS copyright credit line is noted on the appropriate page(s).
>
> Publishing implications of electronic publication of thesis and dissertation material
> Students and their mentors should be aware that posting of thesis and dissertation material on the Web prior to submission of material from that thesis or dissertation to an ACS journal may affect publication in that journal. Whether Web posting is considered prior publication may be evaluated on a case-by-case basis by the journal's editor. If an ACS journal editor considers Web posting to be "prior publication", the paper will not be accepted for publication in that journal. If you intend to submit your unpublished paper to ACS for publication, check with the appropriate editor prior to posting your manuscript electronically.

Reuse/Republication of the Entire Work in Thesis or Collections: Authors may reuse all or part of the Submitted, Accepted or Published Work in a thesis or dissertation that the author writes and is required to submit to satisfy the criteria of degree-granting institutions. Such reuse is permitted subject to the ACS' "Ethical Guidelines to Publication of Chemical Research" (http://pubs.acs.org/page/policy/ethics/index.html): the author should secure written confirmation (via letter or email) from the respective ACS journal editor(s) to avoid potential conflicts with journal prior publication*/embargo policies. Appropriate citation of the Published Work must be made. If the thesis or dissertation to be published is in electronic format, a direct link to the Published Work must also be included using the ACS Articles on Request author-directed link - see http://pubs.acs.org/page/policy/articlesonrequest/index.html

* Prior publication policies of ACS journals are posted on the ACS website at http://pubs.acs.org/page/policy/prior/index.html

If your paper has not yet been published by ACS, please print the following credit line on the first page of your article: "Reproduced (or 'Reproduced in part') with permission from [JOURNAL NAME], in press (or 'submitted for publication'). Unpublished work copyright [CURRENT YEAR] American Chemical Society." Include appropriate information.

If your paper has already been published by ACS and you want to include the text or portions of the text in your thesis/dissertation, please print the ACS copyright credit line on the first page of your article: "Reproduced (or 'Reproduced in part') with permission from [FULL REFERENCE CITATION.] Copyright [YEAR] American Chemical Society." Include appropriate information.

Submission to a Dissertation Distributor: If you plan to submit your thesis to UMI or to another dissertation distributor, you should not include the unpublished ACS paper in your thesis if the thesis will be disseminated electronically, until ACS has published your paper. After publication of the paper by ACS, you may release the entire thesis (**not the individual ACS article by itself**) for electronic dissemination through the distributor; ACS's copyright credit line should be printed on the first page of the ACS paper.

v1.4

Questions? E-mail us at customercare@copyright.com.

2. Comparison of QM Methods for the Evaluation of Halogen−π Interactions for Large-Scale Data Generation

https://pubs.acs.org/doi/full/10.1021/acs.jctc.5c00456

3. A QM-AI Approach for the Acceleration of Accurate Assessments of Halogen-π Interactions by Training Neural Networks

https://pubs.acs.org/doi/10.1021/acs.jcim.5c02136

4. Expanding and Enhancing Neural Network-Based QM-AI Models for the Accurate Prediction of Halogen-π Interaction Energies in Protein Contexts

https://pubs.acs.org/doi/full/10.1021/acs.jcim.5c03249

1. Introduction

1.1 Introduction on Halogen Bonding

A comprehensive understanding of all factors contributing to protein-ligand recognition is essential for modern drug discovery and design. Although extensive research has been conducted on classical molecular interactions such as hydrogen bonding, van der Waals contacts, ion pairs, ion-dipole interactions, cation···π, C-H···π, and π···π interactions, other types of noncovalent interactions, including halogen and chalcogen bonds (generally known as σ-hole interactions), have attracted significant attention over the past decades but still remain not fully understood.[1-8] Halogen bonding (XB) is characterized by the directional attraction between an electrophilic region on a halogen atom (σ-hole), typically chlorine, bromine, or iodine, and a nucleophilic region in another entity, such as the protein binding site.[9-11] Its main appeal lies in its high directionality, which arises from a strongly anisotropic electron distribution around the halogen atom.[12] This anisotropy generates regions of high lateral electron density around the halogen atom perpendicular to the R-X axis, and a corresponding σ-hole, a region of positive electrostatic potential in elongation of the R-X bond axis.[13] Computational methods have been extensively applied to accurately characterize the geometric and energetic features of halogen bonding.[14-18] The strength of halogen bonds generally follows the trend Cl < Br < I, reflecting the increasing size and polarizability of the σ-hole for the heavier halogens. In systems where the substituent R exerts a strong electron-withdrawing effect on the halogen X, pronounced "tuning effects" can further enhance both the strength of the halogen bond and the accessibility of the σ-hole.[19-23] The spatial arrangement of the halogen bond complex also plays a critical role in determining interaction strength. Two key geometric parameters influence this effect: (i) The halogen bond distance and (ii) the σ-hole angle.[24, 25] When the interacting atoms are positioned too closely, repulsive forces dominate, whereas increasing the distance beyond the optimal range leads to a gradual reduction in interaction strength. Similarly, deviations from the optimal σ-hole angle (α = 180°) result in a substantial loss of interaction energy ΔE, with the attractive component essentially disappearing at angular deviations exceeding approximately 40°. In a simplified definition, the interaction energy ΔE is calculated as:

$$\Delta E = E_{XB-complex} - \left(E_{XB-donor} + E_{XB-acceptor}\right) \quad \text{(eq. 1)}$$

where $E_{XB\text{-}complex}$ denotes the system energy of two molecules in complex, and $E_{XB\text{-}donor}$ and $E_{XB\text{-}acceptor}$ denote the individual system energies of the single molecules separately. Beyond these electronic and geometric features, halogen bonds are distinguished by their versatility in interacting with a broad variety of binding-site partners and by their capacity to form complex interaction networks analogous to hydrogen bonds.[26, 27] Owing to this combination of directionality, tunability, and adaptability, halogen bonds have gained considerable importance in medicinal chemistry and drug design, where they can enhance ligand binding affinity and specificity, improve protein-ligand complex stability, and even promote unconventional binding modes that expand the possibilities of molecular recognition.[6-8, 28-42]

1.2 Halogen Bond Acceptors

Halogen bonds exhibit considerable versatility in biological systems, forming with a wide range of electron-rich acceptors, including lone-pair donors such as oxygen, nitrogen, and sulfur atoms, as well as π-systems and even solvent molecules.

To date, many of these nucleophilic interaction partners have been studied systematically. Among these, the most common acceptor is the carbonyl oxygen of the protein backbone, which is abundant throughout biomolecular structures. In a systematic study, Wilcken *et al.*[24, 25] analyzed interactions between halogenated ligands and backbone carbonyl groups in protein crystal structures from the Protein Data Bank (PDB)[43]. Using quantum mechanical (QM) calculations on representative complexes, they determined adduct formation energies and provided insights into the characteristic geometric features governing halogen bond formation. Their results indicated that the most favorable halogen bond geometry features an interaction distance of roughly 3.1 Å and a σ-hole angle of $\alpha = 180°$, underscoring the highly directional nature of halogen bonds. Zimmermann *et al.*[44, 45] extended this previous work by investigating halogen bonds onto the π-system of the peptide bond of *N*-methylacetamide as a surrogate for the protein backbone. Halogen bonds can also form with the sulfur atoms of methionine and cysteine, typically involving the interaction with a lone pair of the sulfur.[46] Interestingly, sulfur itself possesses a σ-hole along the R-S bond axis, enabling it to participate in σ-hole interactions known as chalcogen bonds.[47, 48] Beyond the protein backbone and sulfur-containing residues, other potential halogen bond acceptors have been investigated, including the nitrogen atoms

of the imidazole ring in histidine[49], the carboxylate oxygens of aspartate and glutamate[50], and the carboxamide oxygens and nitrogens of asparagine and glutamine[50]. The hydroxyl groups of serine, threonine, and tyrosine have likewise been proposed as halogen bond acceptors, although systematic studies of these interactions remain limited.[7, 35] Similar to these hydroxyl functionalities, the oxygen atoms of water molecules can also engage in halogen bonding. In 2010, Zhou *et al.*[51] provided valuable insights into this aspect by examining crystal structures of halogenated nucleic acids, proteins, and small molecules, and by introducing the concept of halogen-water-hydrogen (XWH) bridges. Given the ubiquity of water in protein binding sites, its role in mediating protein-ligand interactions is particularly relevant for drug design.[52-60] However, accounting for water molecules in such systems can be challenging. Water molecules located deep within the binding pocket, often referred to as "*interstitial waters"*, can establish extensive hydrogen-bonding networks with both the protein and the ligand. While retaining such waters is entropically unfavorable due to restricted mobility, their presence can be enthalpically advantageous, as they contribute multiple stabilizing interactions.[61, 62] A net increase in binding affinity is achieved only when this enthalpic gain compensates for the entropic loss. Conversely, displacing ordered water molecules, especially those located near hydrophobic surfaces, can enhance binding affinity by releasing them into the bulk solvent and thus increasing the overall system entropy.[63-69]

Another major class of halogen bond acceptors in protein-ligand-interactions are the electron-rich π-systems of aromatic amino acid sidechains of tyrosine, phenylalanine, histidine, and tryptophan, where systematic evaluations are still underrepresented. The π-systems of these residues are characterized by delocalized electron clouds above and below the aromatic ring plane, providing regions of high electron density that can interact favorably with the positively polarized σ-hole of a halogen atom.[36, 70-75] However, accurately describing and quantifying such halogen$\cdots\pi$ interactions remains challenging, as they often coexist and compete with other noncovalent forces, including $\pi\cdots\pi$ stacking, cation$\cdots\pi$, and van der Waals interactions. Moreover, the anisotropic nature of the halogen's electrostatic potential and the spatially diffuse character of the π-electron density make these interactions highly dependent on geometric orientation and local chemical environment. Understanding halogen$\cdots\pi$ interactions therefore requires careful theoretical and computational treatment,

allowing genuine σ-hole interactions to be distinguished from more general or secondary contacts, e.g. π···π or C-H···π contacts.[76] In 2012, Rezac *et al.*[77] conducted a comprehensive benchmarking study to evaluate how different QM methods and basis sets perform in describing various types of noncovalent interactions involving halogenated molecules. From this work, they established a small but valuable benchmark data set of interaction geometries. To specifically assess halogen···π interactions, the authors employed simple model systems such as halomethane and its tuned derivative, trifluorohalomethane, in complex with benzene. In 2020, Zhu *et al.*[78] presented a perspective on the application of QM approaches for investigating halogen bonding, extending the analysis to include aromatic acceptor systems. Despite these important contributions, comprehensive large-scale analyses, particularly those focusing on halogen bond donors and acceptors of relevance to drug discovery remain scarce. Accurate computational modeling of such interactions is crucial for understanding their energetic significance in protein-ligand binding and for their rational use in structure-based drug design. Along similar lines, Wallnöfer *et al.*[79] explored halogen····π interactions using chlorobenzene and bromobenzene in complex with *p*-cresol, offering an initial comparison of different QM methods and basis sets for these systems. These experimental and computational findings highlight both the diversity and complexity of halogen bond acceptor environments. To accurately describe and quantify such interactions, particularly in the context of halogen···π systems, reliable QM methods are required.

1.3 Quantum Mechanical Methods and Basis Sets

To gain a deeper understanding of the underlying forces and directional preferences that define halogen bonding, computational chemistry has become an indispensable tool.[14, 16, 76, 80] Quantum mechanical (QM) calculations allow for a detailed characterization of the electronic features responsible for halogen bond formation, such as the presence of the σ-hole, charge distribution, and polarization effects. These methods enable the decomposition of interaction energies into their electrostatic, dispersion, and induction components, thereby providing insights that are difficult to obtain experimentally. Furthermore, molecular dynamics and hybrid QM/MM (molecular mechanics) approaches extend this understanding to complex biological systems, where environmental effects and conformational flexibility can play crucial

roles.[81] Together, these computational techniques complement experimental observations and offer a molecular-level perspective on the nature, strength, and variability of halogen bonding in chemical and biological contexts. A variety of QM methods exist that differ in how they treat electronic interactions, exchange, and correlation effects, which in turn determine their accuracy and computational cost.[82, 83]

Balancing these factors is particularly important for accurately describing complex biological systems. The Hartree-Fock (HF) method serves as the foundation of most quantum mechanical approaches.[84] The electronic structure of a molecule is described as the average potential of all other electrons in a system, where molecular wavefunctions are expressed as Slater determinants, which inherently account for electron exchange effects exactly. However, HF neglects electron correlation, the instantaneous interactions between electrons, leading to inaccuracies in describing properties such as reaction energies or dispersion forces. To address these limitations, post-Hartree-Fock (post-HF) methods have been developed that explicitly account for electron correlation.[85, 86] Configuration Interaction (CI) methods do so by constructing a linear combination of multiple Slater determinants, improving accuracy but at the cost of computational efficiency.[87]

Møller-Plesset perturbation theory (MPn)[88] represents a class of post-HF methods that incorporate electron correlation through a systematic perturbative expansion of the wavefunction derived from the simple HF reference. Among these, the second-order Møller-Plesset method (MP2)[89] is particularly popular, as it offers a favorable balance between computational cost and accuracy for a wide range of molecular systems. Building upon this, MP3[90, 91] includes third-order corrections, providing a more refined treatment of electron correlation, albeit with a substantially higher computational expense. Recognizing the limitations of both, intermediate variants such as MP2.5[92] have been introduced. MP2.5 effectively averages the MP2 and MP3 energies, exploiting the systematic error compensation between MP2's tendency to underestimate and MP3's tendency to overestimate correlation effects. This pragmatic approach often yields results comparable to the more computationally demanding coupled cluster method with single, double, and perturbative triple excitations (CCSD(T))[83], but at a fraction of the cost. Further refinements have also been developed to enhance the predictive power of MP2. For example, Spin-Component-Scaled MP2 (SCS-MP2)[93, 94] can improve the accuracy of the standard MP2 correlation energy calculation across diverse chemical environments by

applying distinct scaling factors to the parallel and opposite spin components. This mitigates the tendency of conventional MP2 to overestimate the effects of electron correlation.

Coupled cluster (CC) methods provide a more rigorous and systematically improvable approach to treating electron correlation than perturbative methods.[95] In CC theory, the electronic wavefunction is expressed using an exponential approach which considers correlated electron excitations from occupied to virtual orbitals. Among the available approaches, the CCSD(T) method is widely considered as the “gold standard,” offering the highest accuracy among nonempirical methods that remain feasible for systems of practical size.[96] Despite being computationally demanding, the exceptional accuracy and reliability of CC methods make them a benchmark and reference for evaluating and developing other quantum mechanical approaches.[97]

Unlike post-HF methods, which explicitly construct correlated wave functions, density functional theory (DFT) methods describe electronic structures in terms of electron densities.[98-100] This makes DFT a more computationally efficient alternative. In DFT, the complex effects of electron exchange and correlation are incorporated through exchange-correlation functionals, which approximate many-body interactions as a function of local electron density and its derivatives. Various families of functionals have been developed to improve the balance between accuracy and efficiency. The TPSS[101] functional, for instance, is a meta-generalized gradient approximation (meta-GGA[102]) that includes the kinetic energy density as an additional variable beyond the local density and its gradient. This extension enhances accuracy for systems involving weak intermolecular interactions, transition metal complexes, and reaction barriers. B3LYP[103, 104] and M06-2X[105] are two widely used hybrid GGA functionals. They incorporate a fraction of exact HF exchange into the DFT exchange-correlation formulation but differ in the amount of exchange contribution. This hybridization reduces systematic errors associated with self-interaction and often yields reliable thermochemical and structural predictions across diverse chemical systems at reliable computational cost. To further improve the treatment of noncovalent interactions, which are often underestimated by conventional DFT, empirical dispersion corrections such as Grimme’s “D3” are frequently employed.[106] These corrections account for long-range van der Waals forces, significantly enhancing the accuracy of DFT calculations.

In addition to selecting an appropriate QM method, the choice of basis set plays a crucial role in determining the accuracy and computational cost of electronic structure calculations.[107] A basis set defines the mathematical functions used to represent molecular orbitals (or electron density in DFT), and its quality and size directly affects the accuracy of the electronic structure description. Commonly used basis sets include the triple-ζ valence with polarization (def2-TZVPP)[108] and its augmented variant TZVPPD[109], which adds diffuse functions to better describe the electron density. The TZVPP basis set, widely used in both density functional theory (DFT) and post-HF methods, offers an excellent balance between computational efficiency and accuracy by including multiple polarization. Another important family is the correlation-consistent polarized valence X-ζ (cc-pVXZ)[110] basis sets, where X represents the cardinal number (D = double, T = triple, Q = quadruple, etc.). These basis sets are systematically constructed to converge toward the complete basis set (CBS) limit, thereby improving the description of electron correlation effects with increasing cardinality. The augmented versions (aug-cc-pVXZ)[111] include diffuse functions that further enhance the treatment of weak, long-range interactions such as dispersion and hydrogen bonding. This is critical for accurately modelling noncovalent complexes and large biomolecules. Some basis sets contain the suffix “-PP” to the names denoting the inclusion of pseudopotentials (or effective core potentials), which replace the explicit treatment of inner-core electrons for heavier elements such as iodine, thereby reducing computational cost without significantly compromising accuracy. Although the most accurate results would theoretically be obtained from CCSD(T) calculations using a complete basis set, such computations are practically impossible. Instead, basis set extrapolation techniques are commonly employed, allowing energies computed with smaller, typically correlation-consistent basis sets, to be extrapolated toward the CBS limit, providing near-converged results at a fraction of the computational cost.[112] Throughout the thesis, the short term “CCSD(T)/CBS” will be used and always refers to this CBS extrapolation approach.

A further consideration in achieving more accurate interaction energies is the basis set superposition error (BSSE), which arises when the basis functions of one interacting fragment artificially lower the energy of another due to basis set overlap. This effect leads to an overestimation of the binding strength, particularly when using smaller basis sets. The most common way to account for BSSE is the counterpoise correction proposed by Boys and Bernardi[113], in which the energy of each monomer is corrected

by subtracting the contribution arising from the overlap of basis functions in the dimer calculation. While this correction can improve the accuracy of interaction energies, its application remains somewhat controversial, as it may sometimes overcompensate or distort the error.[114] Together, the appropriate choice of QM method and basis set determines the achievable balance between computational efficiency and predictive accuracy. Nevertheless, even with these optimizations, high-level QM calculations remain computationally intensive, motivating the development of more efficient energy estimation schemes.

1.4 Classical Scoring Functions

To overcome the limitations of high-level QM calculations, a variety of scoring functions have been developed to provide rapid estimates of intermolecular interaction energies.[115] These methods simplify the underlying physics to achieve computational efficiency and can generally be categorized into three major classes: force-field-based, empirical, and knowledge-based scoring functions.[116-121]

Force-field-based scoring functions, e.g. Autdock[122, 123] or GOLD[124], originate from classical mechanics and model molecular interactions using predefined potential energy terms. These functions typically include bonded contributions, such as bond stretching, angle bending, and torsional terms, as well as nonbonded interactions like electrostatics and van der Waals forces. The parameters of these force fields are calibrated against experimental measurements or high-level quantum mechanical data. Because of their simplicity and computational efficiency, force-field approaches are extensively applied in molecular dynamics simulations, where rapid energy evaluation is essential.

Empirical scoring functions, e.g. FlexX[125] or Glide[126], build on simplified physical models by introducing parameterized terms that are fitted to reproduce experimental binding affinities or high-level QM reference data. These functions typically combine weighted contributions representing various interaction types, such as hydrogen bonding, hydrophobic contacts, electrostatic interactions, and desolvation effects. The parameters are optimized to best match known data sets, allowing empirical models to provide fast and reasonably accurate energy estimates. However, their transferability is often limited outside the chemical space on which they were parameterized.

Knowledge-based scoring functions, e.g. PMF[127, 128] or DrugScore[129], adopt a statistical approach, deriving effective interaction potentials from the frequency with which atom-atom contacts are observed in large structural databases, such as collections of protein-ligand complexes. The underlying assumption is that frequently observed spatial arrangements correspond to energetically favorable interactions. By translating these statistical preferences into potential energy terms, knowledge-based methods capture recurring patterns of molecular recognition in experimental structures. Although computationally efficient and effective in many docking applications, their accuracy depends strongly on the size, diversity, and quality of the structural data used to construct them.

While these conventional scoring functions enable fast and practical energy estimation, their accuracy and transferability remain limited by the assumptions inherent in their parameterization and reference data. In recent years, machine learning (ML)-based scoring functions have emerged as a new generation of models that aim to overcome these constraints by learning complex relationships directly from high-quality, QM or experimental data sets.

1.5 Machine Learning-derived Scoring Functions

Although traditional scoring functions based on force fields, empirical approaches, and knowledge-based methods have each provided valuable frameworks for estimating intermolecular interaction energies, they inevitably depend on simplifying assumptions and fixed functional forms. These constraints often prevent them from generalizing across diverse chemical systems or accurately capturing the complex, nonlinear relationships inherent in molecular interactions. However, recent advances in computational power, data availability and algorithmic development have opened up new opportunities to overcome these limitations through machine learning (ML) approaches. Unlike traditional scoring functions, which are based on predefined analytical expressions, ML-based scoring functions learn the mapping between molecular features and the target property, e.g. interaction energies, directly from data.[130-133] This allows them to reproduce high-level quantum mechanical (QM) accuracy while maintaining computational efficiency. The origins of ML as a scientific discipline trace back to early research in pattern recognition and artificial intelligence in the mid-20th century. Over the past decades, advances in algorithms, data

availability, and computational power have transformed ML into a powerful framework for predictive modeling across numerous scientific domains.[134-139] Many scoring functions, e.g. KDEEP[140] or OnionNet[141, 142], have already been developed for the prediction of protein-ligand binding affinities. In general, ML methods can be classified into three main categories: (i) supervised learning, where models are trained on labeled data to predict target properties; (ii) unsupervised learning, which identifies hidden structures or correlations within unlabeled data sets; and (iii) reinforcement learning, where algorithms iteratively improve their performance by interacting with an environment and receiving feedback.

Within supervised learning, by far the most common paradigm in molecular modelling, a wide range of algorithms has been applied. Linear and nonlinear regression models, decision trees, random forests (RFs)[143, 144], support vector machines (SVMs)[145], and kernel ridge regression (KRR)[146] have all been used to relate molecular descriptors to properties such as binding energies, dipole moments, or reaction barriers. In recent years, artificial neural networks (ANNs)[147-149] have gained significant prominence owing to their exceptional capacity to model complex, non-linear relationships within data. A major advantage of neural-network-based ML methods lies in their scalability and adaptability, which enable them to effectively handle large and diverse data sets. Once trained, these models can predict molecular properties or interaction energies for a vast range of systems within seconds, offering a radical reduction in computational cost compared to QM calculations. Typically, neural networks (NNs) are provided with molecular descriptors that encode essential geometric or electronic information, such as intermolecular distances, angles, or atomic environment features. Through nonlinear transformations, the model learns to map these descriptors to the corresponding interaction energies, effectively capturing complex relationships that are difficult to describe analytically. In a typical artificial neural network, the architecture is organized into multiple layers of interconnected nodes, or “neurons.” These layers usually include an input layer, one or more hidden layers, and an output layer. The input layer receives molecular descriptors, which are then propagated through the hidden layers where each neuron performs a weighted summation of its inputs followed by a nonlinear transformation known as an activation function. Common activation functions include the rectified linear unit (ReLU, or LeakyReLU), sigmoid, and hyperbolic tangent (tanh), each introducing nonlinearity that enables the network to capture complex dependencies between input features and target properties.[150, 151]

The number of hidden layers and the number of neurons per layer determine the model's capacity or "depth," with deeper architectures often providing enhanced representational power at the cost of increased training complexity and risk of overfitting. During training, the network's weights are iteratively optimized using algorithms such as stochastic gradient descent or its variants (e.g., Adam Optimizer), guided by a loss function that quantifies the deviation between predicted and reference values, e.g. mean-squared error (MSE).[152-154] Through this process, the ANN learns an internal representation of the structure-property relationships encoded in the training features, allowing it to generalize and make rapid, accurate predictions for unseen molecular systems. Integrating QM-derived data sets with NN-based models enables researchers to close the gap between computational efficiency and the accuracy of advanced quantum mechanical methods.

The potential of ML to model noncovalent interactions[155-160], in this case particularly halogen bonding, has already been demonstrated in several recent studies. Shaw *et al.* (2019)[161] introduced one of the earliest applications of statistical modeling to halogen bonding, employing a simple two-parameter regression approach to predict interaction energies within a small data set of halogenated complexes. Despite its simplicity, the model achieved a mean deviation of approximately 2.1 kJ/mol on an unseen test set of 80 systems, illustrating how even basic data-driven models can reproduce QM-level accuracy for specific interaction types. Building upon this foundation, Samuel *et al.* (2023)[162] provided a comprehensive overview of ML approaches applied to halogen bonding, emphasizing their growing importance as tools for deciphering the subtleties of molecular recognition. They highlighted that ML serves as a *"powerful tool for unravelling the intricacies of molecular interactions and guiding the design of functional molecular systems,"*[162] and predicted that progress in this field will accelerate as larger, curated data sets become available and hybrid quantum mechanical-machine learning (QM-AI) frameworks gain prominence. Further illustrating this trend, Devore *et al*. (2024)[163] developed a ML-based strategy to characterize halogen bonding interactions using molecular fingerprints as input descriptors. Their results demonstrated that supervised learning algorithms can not only classify different halogen-bond donor types but also accurately predict corresponding interaction strengths. Collectively, these studies underscore the growing synergy between QM data and ML.

In summary, the integration of ML techniques into computational chemistry represents a powerful advancement toward achieving quantum-level accuracy at a fraction of the computational cost. By leveraging data derived from high-level QM calculations, ML models can capture subtle electronic effects and complex, non-linear dependencies that are often inaccessible to traditional scoring functions. Their flexibility allows them to generalize across diverse molecular systems, while their scalability makes them well-suited for handling large data sets. As research in this field continues to evolve, hybrid QM-ML approaches are expected to play an increasingly central role in modeling noncovalent interactions with both efficiency and precision.

1.6 Goal of the Thesis

The comprehensive goal of this thesis is to advance the understanding and computational modeling of halogen bonding, with a particular focus on its role in biomolecular systems and protein-ligand interactions. While halogen bonding has become an increasingly recognized noncovalent contribution in molecular recognition and drug design, many aspects of its energetic and geometric behavior remain insufficiently explored, particularly in interactions with solvent molecules and aromatic π-systems that are commonly present in protein binding sites and can substantially influence ligand binding and stabilization. To address these gaps, this work follows a two-part structure that reflects both a fundamental and a methodological perspective.

The first part of this work, described in Chapter 1 (Publication 1), presents a quantum mechanical (QM) investigation of halogen···water interactions, focusing on their energetic and structural characteristics in biologically relevant environments such as protein binding sites. Protein crystal structures from the Protein Data Bank (PDB) are analyzed to identify and statistically evaluate halogen···water interactions in biological systems. The extracted interaction geometries are subsequently examined using QM calculations to determine interaction energies and characteristic geometric parameters. Together, these findings provide new insight into solvent-mediated halogen bonding and highlight the relevance of water molecules in stabilizing protein-ligand complexes under specific conditions.

The second, and more extensive part of the thesis, focuses on halogen···π interactions, integrating high-level QM calculations with machine learning (ML) approaches. This work is detailed in Chapters 2 and 3, which together form the methodological core of the thesis. In Chapter 2 (Publication 2), different QM methods and basis sets are systematically benchmarked to identify accurate and computationally efficient protocols for modeling halogen···π interactions in representative systems. The benchmarking results provide a foundation for generating consistent, high-quality data sets that capture the energetic and geometric features of these noncovalent interactions. In Chapter 3 (Publications 3 and 4), these findings are further utilized to generate comprehensive data sets comprising halogen···π interaction geometries. These data sets are employed to train neural network models on high-level QM data, enabling the prediction of interaction energies with near-quantum accuracy at substantially reduced computational cost. The models are evaluated on unseen test

sets comprising both random geometries and structures extracted from the PDB, and their limitations are analyzed in relation to key geometric features. Furthermore, the expandability of NN-models is demonstrated by retraining the models on additional data to enhance the predictive performance and diminish outliers. The resulting models not only demonstrate the potential of ML to accelerate high-accuracy interaction energy predictions but can also be integrated into docking or scoring frameworks to improve the assessment of halogen-bonding contributions in protein-ligand systems. Together, these studies advance both the understanding and computational modeling of halogen bonding in biological systems.

2. Results and Discussion

Chapter 1 – Publication 1: Halogen Bonding on Water - A Drop in the Ocean?

Upon solving the crystal structure of the Dual-specificity tyrosine phosphorylation-regulated kinase 1A (DYRK1A)[164], an important regulator in proliferation and differentiation of nerve cells, in complex with the inhibitor 2-cyclopentyl-7-iodo-1*H*-indole-3-carbonitrile[165] (PDB: 8R8E[166]), an interesting binding mode characterized by a halogen···water interaction was observed. Specifically, the ligand's iodine atom engages with a nearby "interstitial" water molecule, forming part of an extended network of halogen and hydrogen bonds within the active site. The presence of this water molecule in the immediate vicinity of the iodine atom suggests the formation of a potential halogen bond. The measured interaction distance ($d_{I\cdots O}$ = 3.35 Å) and σ-hole angle ($\alpha_{C\text{-}I\cdots O}$ = 168.3°) between the iodine and the water oxygen closely match the geometrical parameters previously reported for favorable halogen bond complexes. Although said interaction could only be observed in chain A of the protein, examination of the electron density strongly supports the existence of the water molecule. For a detailed analysis, a section of the binding site including the ligand, the addressed water molecule, and all surrounding amino acid residues within 4 Å was extracted, protonated using the Protein Preparation Wizard within the Schrödinger Suite[167], and subsequently geometrically optimized at a TPSS-D3/TZVP[108] level of theory using TURBOMOLE[168]. To maintain the experimentally obtained binding conformation, heavy atoms were kept frozen during this process, thus only optimizing the hydrogen bond network. From the optimized structure (Figure 1[a]), geometries of the interacting partners were extracted, and adduct formation energies were obtained from MP2/TZVPP single-point calculations of the isolated components. The halogen bond between the ligand's iodine and the water's oxygen atom (Figure 1a, A) exhibited an interaction energy of ΔE = -8.44 kJ/mol for the original ligand. Substituting the ligand with the model compound iodobenzene resulted in a comparable value of ΔE = -8.54 kJ/mol, aligning with previously reported data for non-tuned haloaryl systems, though slightly weaker in magnitude. The reported values represent pairwise adduct formation energies and do not account for cooperative and synergistic effects that may occur within the interaction network. These observations

[a] Figure and table references refer to figures and tables from the corresponding papers and are reproduced in the Appendix.

prompted us to perform a more systematic analysis of such halogen···water interactions across the PDB to evaluate their frequency, geometric preferences, and potential contribution to ligand binding and stability.

A total of 197,120 crystal structures were analyzed with 7635 unique structures containing ligands that bear chlorine, bromine, or iodine moieties. Fluorine contacts were neglected because fluorine naturally does not possess a σ-hole. A total of 3780 water contacts in 1726 unique PDB structures were found. The majority of contacts is attributed to chlorine followed by bromine, and iodine (Table 2). We analyzed interaction distance ($d_{X\cdots O}$) and σ-hole angle ($\alpha_{C\text{-}X\cdots O}$) and angle distributions of all contacts (Figure 2), where distances below 2.0 Å were neglected indicating either clashes or artifacts in the crystallography process. Chlorine and bromine distributions show densed regions of 3.5 Å to 3.9 Å with interaction angles of 80° to 130°. The iodine distribution shows a slightly narrowed region of 3.5 Å to 3.7 Å and angle values of 90° to 110°. To focus only on relevant σ-hole interactions, geometric constraints were applied with an interaction distance of <4 Å and an interaction angle of >140°. A total of 741 examples featuring a promising XB interaction were subsequently examined. B-factors and electron density of the ligands and the interacting water molecules were inspected resulting in a decreased data set of 721 structures. Binding site sections of these structures were extracted containing the ligand, water molecules, as well as surrounding amino acid residues within 4 Å. Ligand analysis across PDB crystal structures shows substantial variation in size and scaffolds, with halogenated aromatic or heteroaromatic scaffolds commonly observed. For simplicity and comparability, each ligand was replaced by the corresponding halobenzene model, which is exactly matched onto the halogen and ring system of the original ligand. Subsequently, extracted binding site sections were protonated, followed by a hydrogen bond network optimization on a TPSS-D3/TZVP level of theory. To avoid overrepresentation of interaction geometries, only unique interactions within a PDB structure were considered. Interactions that occur throughout multiple chains of the same crystal structure were neglected, resulting in a final data set of 516 unique structures (386 chlorine, 104 bromine, and 26 iodine interactions).

To evaluate adduct formation energies and quantify halogen bond strengths, MP2/TZVPP single-point calculations were performed on complexes composed of halobenzene and interacting water. On average, chlorine and bromine interactions

exhibited comparable mean energies while iodine interactions were notably stronger (Table 3). The observed increase in interaction strength from chlorine to iodine reflects the growing polarizability and van der Waals radii of the halogens, consistent with the larger σ-hole character of the heavier atoms. The geometric diversity of these interactions was further analyzed by plotting the interaction distance against the interaction angle (Figure 4). Chlorine interactions (N = 341) displayed a wide distribution of moderately stabilizing energies, while bromine interactions (N = 88) exhibited a similar spatial pattern but slightly enhanced energies. Iodine interactions, though less frequent (N = 25), demonstrated the strongest adduct formation energies and a clear trend of increasing stability with larger σ-hole angles. Among the halogens, iodine exhibited the strongest interaction (ΔE = -9.83 kJ/mol) with an interaction distance of $d_{I\cdots O}$ = 3.25 Å and an angle of $\alpha_{C\text{-}I\cdots O}$ = 175.5°, followed by bromine (ΔE = -6.34 kJ/mol, $d_{Br\cdots O}$ = 3.62 Å, $\alpha_{C\text{-}Br\cdots O}$ = 142.3°) and chlorine (ΔE = -6.02 kJ/mol, $d_{Cl\cdots O}$ = 3.52 Å, $\alpha_{C\text{-}Cl\cdots O}$ = 142.8°). The corresponding geometries were examined, showing a classical σ-hole interaction of the iodine perpendicular to the plane formed by the water atoms, thus addressing the oxygen's lone pair. Chlorine and bromine exhibit a "shifted, anti-parallel" interaction pattern relative to the O-H bond vector of the water molecule, indicating a concurrent involvement of halogen bonding and hydrogen bonding. In this configuration, the halogen atom acts as a halogen bond donor toward the oxygen atom while concurrently serving as a hydrogen bond acceptor through its negatively charged equatorial belt (Figure 6 and 7). Therefore, both directionality and spatial orientation of the halobenzene relative to the water molecule's lone pairs are critical for stabilizing σ-hole-mediated interactions. Accurate assessment of these features requires optimization of the water molecule's proton positions (or ideally the whole hydrogen bond network) within the binding site. Investigating the energy-dependence of this spatial distribution (Figure 8) reveals that chlorine atoms are positioned closer to the hydrogen atoms, while bromine shifts slightly above and iodine distinctly toward the oxygen atom. This directional trend becomes less pronounced in weaker interactions. Additionally, we examined the spatial distribution of the C-X bond vectors addressing the water molecule (Figure 9) to identify directional preferences of the respective halogen interactions. Iodine interactions predominantly exhibit a perpendicular σ-hole orientation toward the water molecule plane, directly addressing the oxygen lone pairs, though some also adopt the "shifted, anti-parallel" geometry, but with significantly weaker interaction energies. Bromine interactions largely maintain

σ-hole directionality toward the lone pairs, with occasional deviations toward the shifted orientation. In contrast, chlorine interactions primarily display this mixed “shifted, anti-parallel” pattern, reflecting a combined halogen- and hydrogen-bonding character. This tendency likely arises from chlorine’s smaller σ-hole and more pronounced lateral electron density.

To address the imbalance in the experimental data set containing predominantly chlorinated ligands, additional virtual complexes were generated using a matched molecular pair approach. In this halogen-centered procedure, both halogen atoms were precisely superimposed, followed by alignment of the C-X bond and the aromatic scaffold, thereby simulating halogen exchange while preserving the original binding geometry. This strategy enables evaluation of how substitution with different halogens may enhance or reduce interaction energies without altering the overall binding mode. Within the virtually enhanced data set, previously found trends become more evident (Figure 12 and 13). Chlorine interactions preferentially adopt shifted antiparallel orientations, engaging partly in halogen and hydrogen bonding. Bromine interactions are positioned slightly above the water oxygen, while iodine atoms localize directly above it, aligning with the lone pairs. Correspondingly, C-Br and C-I bond vectors exhibit increasing σ-hole directionality from bromine to iodine, confirming progressive strengthening of halogen bonding with the water oxygen atom.

Conclusion:

This study demonstrates that halogen···water interactions can contribute favorably to protein-ligand binding, although their overall energetic impact remains moderate. Analysis of the DYRK1A complex with 2-cyclopentyl-7-iodo-1*H*-indole-3-carbonitrile, supported by a systematic PDB survey and QM calculations, revealed versatile halogen···water geometries depending on halogen type and local environment. The strongest interaction observed (≈-10 kJ/mol) is notably weaker than typical halogen bonds to protein residues such as backbone carbonyls (≈-20 kJ/mol)[25] or carboxylates (up to -56 kJ/mol)[50]. While these weaker interactions may seem marginal, their modest strength likely reflects a balance with the desolvation cost of water molecules, suggesting limited but potentially exploitable contributions to molecular recognition.

In addition to enthalpic contributions, the observed halogen···water interactions likely involve enthalpy-entropy compensation. Chlorine···water interactions, characterized by more flexible and “shifted, antiparallel” geometries, may benefit from greater entropic freedom despite weaker interaction energies. In contrast, iodine···water interactions exhibit strong directionality and higher enthalpic stabilization but reduced conformational flexibility, leading to a lower entropic contribution. These trends suggest that the balance between enthalpy and entropy depends strongly on the halogen type, with chlorine favoring flexibility and iodine favoring specificity in binding.

Halogen bonding to interstitial water molecules, those stably integrated into the binding site through multiple hydrogen bonds, may offer an enthalpic advantage and even contribute to ligand selectivity. The extent of this effect depends strongly on the specific spatial context of the water molecule within the binding site. Consequently, systematic investigations are needed to assess when such water molecules represent meaningful targets in ligand design.

Chapter 2 – Publication 2: Comparison of QM Methods for the Evaluation of Halogen-π Interactions for Large-Scale Data Generation

The accurate description of halogen bonding and related noncovalent interactions remains a topic of active research within computational chemistry. Various studies have reported differing conclusions regarding the suitability of quantum mechanical (QM) methods for modeling halogen bonds, reflecting the delicate balance between computational efficiency and accuracy required for these weak interactions. Building on previous findings, where the MP2/TZVPP level of theory proved effective for modeling halogen bonds, this chapter investigates whether the same approach can reliably describe halogen···π interactions. In addition, alternative QM methods, including M06-2X, are evaluated to determine whether more suitable options exist for this interaction type. The comparison between MP2 and M06-2X was of high interest because M06-2X showed very accurate results particularly for interactions with dispersion contributions (including XBs). Therefore, a series of benchmark calculations on different combinations of QM methods and basis sets was performed to evaluate their accuracy and suitability.

We systematically generated interaction geometries of iodobenzene in complex with benzene (as XB acceptor). Iodobenzenes were placed on a regular grid of data points in five different distances (defined as the perpendicular distance between the halogen atom and the benzene plane), and with an optimal σ-hole angle of $\alpha_{\text{C-I}\cdots\pi\text{-plane}}$ = 180°. Adduct formation energies ΔE of the individual interaction geometries were calculated. Results were compared to reference calculation on a CCSD(T) level of theory, extrapolated to the basis set limit using the approach proposed by Halkier *et al*.[112] Due to extraordinarily high computational cost of these calculations, a smaller subset of ~30% of all geometries was used. Mean energy deviations (ΔΔE), mean absolute energy deviations (|ΔΔE|), and root-mean-square errors (RMSE) in kJ/mol from the reference CCSD(T)/CBS, as well as the corresponding computational cost in CPU hours were investigated (Table 1). Differences are calculated as:

$$\Delta\Delta\mathrm{E} = \Delta\mathrm{E}_{\mathrm{method}} - \Delta\mathrm{E}_{\mathrm{CCSD(T)}/CBS} \qquad \text{(eq. 2)}$$

Mean differences indicate a method's overall tendency to over- or underestimate energies but can be misleading. Additionally, mean absolute energy deviations and

RMSD are reported, giving further insights into the magnitude of differences. The runtime is averaged over all single-point calculations.

With a mean energy difference of $\Delta\Delta E$ = -0.23 kJ/mol, a mean absolute difference of $|\Delta\Delta E|$ = 0.64 kJ/mol, an RMSD = 0.91 kJ/mol, and a notably low computational cost of ~1.16 CPU hours, MP2/TZVPP shows an excellent balance between accuracy and computational cost (Figure 1). Applying BSSE correction at the MP2/TZVPP level led to slightly reduced accuracy and a notable increase in computational time. Similarly, the diffuse-function-enhanced TZVPPD basis set yielded the poorest performance among the triple-ζ variants, further confirming that the addition of diffuse functions does not necessarily improve accuracy for these systems. Among the correlation-consistent basis sets, cc-pVTZ-PP delivers excellent accuracy ($\Delta\Delta E$ = 0.59 kJ/mol, RMSD = 0.73 kJ/mol) at a total computational cost of ~1.23 CPU hours, representing only a modest overhead relative to TZVPP and making it a practical and efficient alternative. In contrast, larger and augmented variants such as cc-pVQZ-PP and aug-cc-pVXZ-PP offer no significant improvement in accuracy but require drastically longer runtimes, rendering them impractical for large-scale applications. Although the MP3 method is expected to provide improved accuracy over MP2, the results show substantially larger deviations while requiring significantly more computational time, making it impractical for routine use. The intermediate MP2.5 approach, which averages MP2 and MP3 energies to balance their opposing errors, likewise offers no improvement in accuracy, as its computational demand remains dominated by the costly MP3 calculations.

Among the tested DFT methods, B3LYP(-D3) and TPSS(-D3) show noticeably larger deviations from the reference data, and the inclusion of BSSE correction further increases these discrepancies. Despite its computational efficiency (~0.5 to 4 CPU hours depending on the basis set), M06-2X consistently yields larger deviations than MP2, with errors of several kJ/mol even for medium-sized basis sets. BSSE correction and the addition of diffuse functions both worsen the agreement, and M06-2X/TZVPPD produces the highest RMSD of all tested methods. Overall, DFT methods reproduce general energetic trends but lack the quantitative accuracy required for reliable predictions of halogen$\cdots\pi$ interaction energies. A detailed grid-based analysis was performed for the five halogen$\cdots\pi$ distances to visualize mean adduct formation energies (ΔE) and deviations from the CCSD(T)/CBS

reference (ΔΔE). “2D energy surface plots” revealed consistent interaction trends across all distances, with both MP2/TZVPP and M06-2X/TZVPP showing the strongest attraction near 3.5 Å (Figure 2). While larger deviations were observed for highly repulsive regions, particularly for M06-2X, these discrepancies may be of limited practical relevance, as they occur outside the energetically favorable interaction range. These analyses further confirm the robustness of MP2/TZVPP across different interaction distances and support its use as a reference-level method for subsequent benchmarking.

We primarily focused on iodine···π interactions, as iodine’s large and highly polarizable σ-hole enables the formation of particularly strong and directional halogen bonds, making it especially relevant in medicinal chemistry. Since the fundamental nature of halogen bonding is consistent across the halogen series, interaction strengths observed for iodine can generally be extended to predict proportionally weaker interactions for bromine and chlorine. However, to ensure comparability and broader applicability of important findings, bromine and chlorine were also included for reference. Accordingly, single-point calculations for chlorobenzene and bromobenzene complexes with benzene were performed only at MP2, MP2 with BSSE correction, and M06-2X levels using the TZVPP basis set. The same set of geometries was employed across all halogens to maintain consistent halogen···π distances and enable a direct comparison of interaction trends. For chlorine and bromine, the overall interaction trends closely mirror those observed for iodine. MP2/TZVPP provides consistently good agreement with the CCSD(T)/CBS reference at minimal computational cost (~1-1.3 CPU hours), making it an efficient choice for these halogens as well. Application of BSSE correction slightly improves accuracy for chlorine and bromine but at more than triple the computational time, reducing its applicability for large-scale calculations. In contrast, M06-2X(-D3) reproduces similar qualitative trends but yields roughly twice the energy deviations of MP2, indicating that, despite their low computational cost, these methods are less reliable for quantitative energy predictions.

Conclusion:

We demonstrated that MP2/TZVPP provides an excellent balance between accuracy and computational efficiency for describing halogen···π interactions, showing close agreement with CCSD(T)/CBS reference data. These findings indicate that MP2

captures the essential properties of these noncovalent interactions, while offering a substantial computational speed-up of approximately two orders of magnitude (10^2) compared to CCSD(T). This performance is consistent across iodine, bromine, and chlorine, confirming the method's reliability for different halogen types. In contrast, commonly used DFT functionals, including M06-2X, failed to achieve comparable quantitative accuracy and are therefore less suitable for precise energy evaluations in this context. While BSSE correction can slightly improve accuracy for lighter halogens, its computational cost limits practicality for large-scale studies. We use these findings as the basis for generating large-scale, high-quality data sets, which are subsequently employed in machine learning approaches to efficiently model and predict halogen···π interaction energies.

Chapter 3.1 – Publication 3: A QM-AI Approach for the Acceleration of Accurate Assessments of Halogen-π Interactions by Training Neural Networks

Building on our previous investigations of quantum mechanical (QM) methods suitable for describing halogen···π interactions, we now extend those findings toward their integration into machine learning models (neural networks) through a systematic analysis of halogen···π complexes. With the goal of developing NN-models for integration into the drug design process, such as virtual screening and molecular docking, we focus on halogen···π interactions in the context of protein-ligand complexes. These interactions typically occur between the halogen atoms of the ligand and the aromatic side chains of residues such as phenylalanine, tyrosine, histidine, and tryptophan. To systematically generate representative interaction geometries, we employed halobenzenes (chlorobenzene, bromobenzene, and iodobenzene) in complex with a benzene molecule, which serves as a simplified model of the phenylalanine side chain. For each halobenzene-benzene complex, interaction geometries were generated by positioning the halogen atom at fixed halogen···π-plane separations ranging from 2.75 Å to 4.50 Å. The orientation of the C-X bond was constrained to deviate by no more than 40° from the normal of the aromatic plane (Figure 7, Materials and Methods), ensuring that the resulting data set primarily captures geometries characteristic of σ-hole-driven halogen···π interactions while minimizing secondary effects such as π···π or C-H···π contacts. Almost 1.4 million single-point calculations were conducted on an MP2/TZVPP level of theory to obtain adduct formation energies. Geometric descriptors comprising pairwise distances and angular parameters were derived from all interaction geometries, ensuring invariance with respect to translation and rotation of the coordinate system. The resulting data set was partitioned into distinct training, validation, and test subsets. Model optimization involved an extensive hyperparameter and architecture search conducted within a leave-one-out five-fold cross-validation framework. The training subset served to fit the neural network weights, while the validation set was used to monitor convergence and refine model parameters. Final predictive performance was evaluated on an independent test set (Figure 8, Materials and Methods). The models were trained in a supervised learning scheme using the mean-squared error as the loss function. The final model demonstrated excellent predictive performance, achieving an $R^2 = 0.9979$ and maintaining low RMSE values across both validation and test sets, indicating consistent accuracy and strong generalization to unseen data (Figure 1 and 2).

To assess the model's generalization beyond the training distribution, a new data set was generated using single-point calculations on interaction geometries with randomly varied translational and rotational features. The model demonstrated strong predictive performance on this unseen data, achieving an $R^2 = 0.856$ and an RMSE of 1.33 kJ/mol. However, several outliers with large energy deviations were identified. Causes for such behavior generally originates either from data points lying outside the training feature space, indicating limited extrapolation ability, or from inherent model limitations such as overfitting or insufficient representational capacity. To further investigate these deviations, the Mahalanobis distance (MD) was employed as a reliability measure to quantify how far individual data points statistically deviate from the training feature distribution. Higher MD values indicate greater dissimilarity from the training data and may be associated with reduced predictive accuracy, suggesting a reasonable correlation between model reliability and feature-space proximity (Figure 3). Outlier analysis was further illustrated using six representative cases exhibiting large energy deviations (either over- or underestimation of the model), for which the corresponding geometric features were examined in detail (Figure 3, Figure 4, and Table 1). Analysis of these outliers revealed two primary sources of deviation. Some correspond to compact, tilted geometries where the halogen lies very close to the π-plane and near the boundaries of the model's feature space. In these cases, additional stabilizing contributions may arise from π···π contacts and lateral C-H···X interactions, which were insufficiently represented in the training data and thus led to systematic underestimation of the interaction energies. Others involve configurations dominated by close C-H···π contacts, where the model tended to overestimate interaction strengths due to limited exposure to repulsive geometries during training. Overall, these cases exhibit high MD values, confirming that they fall outside the distribution of the training data. The deviations therefore arise mainly from insufficient sampling of extreme geometries, either strongly attractive or repulsive.

To assess real-world relevance, a large-scale PDB scan was performed focusing on protein-ligand complexes involving halogenated ligands. Focusing on phenylalanine residues as the primary aromatic acceptor, a data set composed of 1114 interaction geometries that fulfilled the defined geometric and angular criteria was extracted (Table 2). In a matched molecular pair approach, each ligand was replaced by a simplified halobenzene model to isolate the halogen···π interaction. The halogen atom and its C-X vector were precisely aligned with those of the original ligand, and the

benzene ring was optimally oriented to match the ligand's aromatic system. The phenylalanine side chain was truncated and replaced with a protonated benzene model for consistency. While this approach lacks the chemical versatility of full ligands, it enables a controlled comparison of intrinsic interaction geometries and energies. Single-point calculations on an MP2/TZVPP level were performed using TURBOMOLE and adduct formation energies reported. After excluding two sterically clashing cases, 1112 complexes were analyzed. The model achieved good predictive accuracy with an $R^2 = 0.805$ and an RMSE of 1.57 kJ/mol, indicating reliable transferability to real-world data. Most predictions closely matched the calculated energies, while a few outliers exhibited larger deviations that followed the same patterns observed in the earlier test set arising from geometries at or beyond the training feature space (Figure 5, Figure 6, and Table 3). These include non-ideal orientations, excessively short contacts, and interaction types differing from σ-hole-driven halogen···π interactions. It can be argued, that the model's performance on repulsive geometries is of minor relevance, since such cases are typically excluded by earlier filtering or separate repulsive terms.

To avoid bias introduced by the simplified halobenzene model systems, the original heteroaromatic ligand ring geometries were also analyzed. The resulting predictions showed only minimal deviations (~0.3 kJ/mol), confirming the model's good transferability and robustness across different aromatic ring types.

Regarding accuracy and computational cost, the developed neural network models reproduce MP2-level interaction energies with near-quantum accuracy while reducing computational cost by approximately eight orders of magnitude (10^8) compared to the underlying ab initio calculations. This immense acceleration enables the large-scale application of accurate halogen···π interaction scoring in molecular modeling and drug design contexts.

Conclusion:

In this study, we developed neural network models to predict the interaction energies of halogen···π complexes between halobenzenes and benzene with near-quantum accuracy. Trained on geometric descriptors extensive MP2/TZVPP data sets, the models achieved excellent predictive performance and maintained strong accuracy

across independent and PDB-derived test sets. Larger errors were limited to geometries outside the training distribution, emphasizing the model's defined applicability to σ-hole-driven interactions. By leveraging a methodological, hierarchical "double jump" from CCSD(T) → MP2 → NN, this QM-AI approach preserves near-CCSD(T) accuracy while achieving speed-ups of up to eight orders of magnitude (10^8) compared to MP2 calculations. In this case, although the approach was specifically designed to describe σ-hole interactions, it can be easily extended by incorporating additional data to enrich the training feature space and capture a broader spectrum of interaction types. Moreover, it can be extended to other halogen bond acceptors, such as the aromatic side chain residues of histidine, tyrosine, and tryptophan.

Chapter 3.2 – Publication 4: Expanding and Enhancing Neural Network-Based QM-AI Models for the Accurate Prediction of Halogen-π Interaction Energies in Protein Contexts

Building on the excellent performance of the previously developed QM-AI models for the accurate description of halogen···π interactions in halobenzene-benzene systems, we now extend this approach towards a broader set of aromatic acceptors. In this follow-up study, halogen···π interactions involving phenol, imidazole, and indole are systematically investigated, representing the aromatic side chains of tyrosine, histidine, and tryptophan, respectively. Several million interaction geometries of halobenzenes complexed with the respective acceptor systems (phenol, imidazole, and indole) were systematically generated and subsequently subjected to single-point calculations at the MP2/TZVPP level. Halogen atoms were positioned at fixed distances ranging from 2.75 Å to 4.50 Å, where a plane of data points was used at each distance (Figure 7, Materials and Methods). The orientation of the C-X bond was constrained to deviate by no more than 40° from the normal of the aromatic plane. To account for the structural and electronic diversity of the three aromatic acceptor systems, the generation of interaction geometries was adapted individually for each case. For phenol, the presence of the hydroxyl substituent introduces an additional interaction site via the O-H group, and especially the lone pairs of the oxygen atom. Accordingly, a complementary set of geometries was generated to specifically probe interactions involving the hydrogen and oxygen atoms of the hydroxyl group, ensuring adequate sampling of σ-hole configurations. In the case of imidazole, the non-protonated nitrogen atom provides an alternative acceptor and lone-pair interaction site. To capture this behavior, additional geometries were constructed with the halogen atom positioned in a hemisphere around the center of the nitrogen, thereby extending the sampling space toward orientations characteristic of halogen···N interactions. For indole, the fused bicyclic ring system required an enlarged sampling volume to comprehensively describe both the five- and six-membered ring surfaces. Consequently, the grid dimensions were significantly expanded, leading to a substantially larger number of generated geometries. Geometric descriptors capturing all relevant pairwise distances and angular relations were derived from the generated interaction geometries, ensuring full independence to translation and rotation of the underlying coordinate system. The data set was divided into training, validation, and test subsets following the same protocol as described previously in chapter 3.1. Model

optimization involved a comprehensive hyperparameter and architecture search performed through leave-one-out five-fold cross-validation. The network was trained in a supervised learning framework using the mean-squared error as loss function, with the validation set used to monitor convergence and tune model parameters, and the independent test set employed to assess final predictive performance (Figure S2). For each acceptor system, an individual model was generated and trained independently. All three models exhibited excellent agreement with reference data, achieving values of R^2_{phenol} = 0.9983, $R^2_{imidazole}$ = 0.9919, and R^2_{indole} = 0.9982, along with consistently low RMSE values ($RMSE_{phenol}$ = ~0.15 kJ/mol, $RMSE_{imidazole}$ = ~0.28 kJ/mol, $RMSE_{indole}$ = ~0.17 kJ/mol) for both validation and test set. These findings confirm the models' reliability and strong generalization performance even on previously unseen geometries.

To circumvent potential bias arising from feature-space overlap between training and test data, model generalization was further evaluated using a separate set of randomly generated geometries. This data set comprised both fully out-of-distribution samples and configurations that lie within the training distribution but correspond to interpolated combinations of geometric parameters. In the first step (i) a subset containing only data points with features lying within the training distribution (2.75 Å $\leq d_{X\cdots\pi\text{-plane}} \leq$ 4.5 Å; $\alpha_{C\text{-}X\cdots\perp(\pi\text{-plane})} \leq 40°$) was analyzed. In the second step (ii) distance and angle constraints were dropped and the full test data set was evaluated. Within the constrained subset (step i), the models effectively reproduced the excellent performance observed during validation, yielding coefficients of determination of $R^2_{phenol(i)}$ = 0.9935, $R^2_{imidazole(i)}$ = 0.9848, and $R^2_{indole(i)}$ = 0.9953, together with RMSE values of 0.27, 0.34, and 0.28 kJ/mol, respectively. These results demonstrate the models' ability to interpolate smoothly within the training distribution. When evaluated on the full test data set (step ii; Figure 1), predictive performance remained strong, with values of $R^2_{phenol(ii)}$ = 0.9644, $R^2_{imidazole(ii)}$ = 0.9096, and $R^2_{indole(ii)}$ = 0.9269, along with RMSE values of 0.90, 0.94, and 1.17 kJ/mol for phenol, imidazole, and indole, respectively.

Outliers exhibiting large deviations between predicted and reference interaction energies were analyzed in detail (Figure 2, Table 1). Most of these correspond to geometries featuring additional stabilizing contributions from $\pi\cdots\pi$ or lateral C-H$\cdots$X interactions, which were underrepresented in the training data and therefore often underestimated by the models. In contrast, a smaller number of outliers originated from

configurations with excessively short halogen···π-plane distances, leading to strong repulsive forces and overestimated interaction strengths. In both cases, the corresponding high Mahalanobis distance (MD) values indicate that these geometries lie outside the feature space covered during training.

Conducting a PDB scan yielded real-world examples from protein crystal structures which were used as an additional evaluation set. Ligands were replaced by a simplified halobenzene model to isolate the halogen···π interaction. The halogen atom and its C-X vector were precisely aligned with those of the original ligand, while the benzene ring was oriented to reproduce the aromatic geometry of the ligand. Each aromatic residue was substituted by its corresponding protonated, QM-optimized model system. The resulting PDB-derived data sets are hereafter referred to as the “tyrosine (TYR)”, “histidine (HIS)”, and “tryptophan (TRP)” sets. These sets comprise 1152 geometries for tyrosine, 661 for histidine, and 384 for tryptophan. Single-point MP2/TZVPP calculations were performed to obtain adduct formation energies. Geometries with interaction energies exceeding +10 kJ/mol were excluded, yielding final data sets of N_{TYR} = 1142, N_{HIS} = 660, and N_{TRP} = 384. The same two-step evaluation protocol was applied.

For the constrained subset (step i), the models again demonstrated excellent predictive performance, achieving R^2 values of 0.9904, 0.9834, and 0.9900 with RMSE values of 0.30, 0.43, and 0.33 kJ/mol for tyrosine, histidine, and tryptophan, respectively. Evaluation of the full PDB-derived data sets (step ii; Figure 3) yielded slightly reduced but still strong performances, with R^2 values of 0.9530 for tyrosine, 0.8840 for histidine, and 0.8819 for tryptophan with RMSE values of 0.76, 1.65, and 1.33 kJ/mol, respectively. As observed previously, most predictions closely matched the reference interaction energies, while a small number of outliers displayed larger deviations consistent with those identified in the randomly generated test sets (Figure 4, Table 3). These deviations primarily originated from geometries located at or beyond the boundaries of the training feature space.

To assess the models’ behavior and performance following data set expansion and retraining, additional data sets comprising 300,000 interaction geometries between halobenzenes (100,000 per halogen) and the three model systems were generated by randomly varying translational and rotational features. All model architectures and hyperparameters were kept unchanged, and each system - phenol, imidazole, and

indole - was retrained using the expanded data set. As a result of the increased data diversity, a slight reduction in training and validation accuracy was observed, with R^2 values of 0.9958, 0.9827, and 0.9952 and corresponding RMSE values of 0.34, 0.45, and 0.39 kJ/mol for phenol, imidazole, and indole, respectively. Importantly, this modest decrease in in-sample performance was accompanied by a notable improvement in predictive accuracy on external data. Evaluation on the randomly generated geometry test set (Figure 5) yielded R^2 values of 0.9851, 0.9555, and 0.9803 with RMSE values of 0.58, 0.86, and 0.80 kJ/mol for phenol, imidazole, and indole, respectively. Consistent with this trend, performance on the PDB-derived test set (Figure 6) also improved, achieving R^2 values of 0.9610for tyrosine, 0.9252 for histidine, and 0.9179 for tryptophan with corresponding RMSE values of 0.69, 1.35, and 1.11 kJ/mol, respectively, further confirming the enhanced generalization of the retrained models.

All three neural network models reproduce MP2-level interaction energies with near-quantum accuracy while achieving a drastic reduction in computational cost. Consistent with our previous findings, the present models likewise provide an acceleration of approximately eight orders of magnitude ($\sim 10^8$) compared to the underlying ab initio calculations.

Conclusion:

In this study, the previously developed QM-AI framework was extended beyond the halobenzene-benzene model to encompass halogen$\cdots\pi$ interactions with phenol, imidazole, and indole, representing the aromatic side chains of tyrosine, histidine, and tryptophan, respectively. Individually trained neural network models reproduced MP2/TZVPP interaction energies with near-quantum accuracy and demonstrated excellent predictive performance across independent, randomly generated, and PDB-derived test sets. Model generalization was assessed in detail using a two-step evaluation strategy that distinguishes between interpolation within the training distribution and extrapolation beyond its boundaries. Within the constrained feature space, the models demonstrated consistently high accuracy, confirming their ability to smoothly interpolate between sampled interaction geometries. When evaluated on fully unconstrained test sets, predictive performance remained robust, with larger deviations confined to a small number of geometries located outside the training

feature space. Larger deviations were limited to geometries lying outside the training feature space, confirming the models' well-defined applicability to σ-hole-driven halogen···π interactions. Furthermore, we demonstrated and evaluated the models' extendibility by incorporating additional data and re-training the initial network which led to improved predictive performance and confirmed the capability to generalize and adapt to an expanded interaction space. Consistent with previous findings, the developed models maintain near-quantum accuracy while reducing computational cost by approximately eight orders of magnitude ($\sim 10^8$) compared to ab initio reference methods.

3. Conclusion and Outlook

The studies presented in this thesis collectively deepen our understanding of halogen-mediated noncovalent interactions and demonstrate how combining high-level quantum mechanics (QM) with machine learning can provide both physical insight and computational efficiency for molecular modeling and drug discovery.

The work began with an exploratory investigation of halogen···water interactions (Publication 1), which served as a stand-alone study and the conceptual foundation for subsequent research. Analysis of a halogenated inhibitor bound to DYRK1A revealed an unusual iodine···water contact that inspired a comprehensive database and QM survey. This systematic evaluation established that halogen···water interactions, although energetically modest, are structurally well defined and exhibit clear trends across the series of halogens. Chlorine typically forms more flexible, mixed halogen-hydrogen-bonding arrangements, whereas iodine engages in highly directional σ-hole interactions with water oxygen lone pairs. These findings demonstrated that interstitial water molecules can participate in subtle but potentially meaningful recognition patterns within protein binding sites. Importantly, this study underscored the sensitive balance between enthalpic stabilization and entropic cost in such systems.

Publications 2 to 4 form the main body of the thesis and focus on halogen···π interactions, which play a central role in biomolecular recognition processes involving aromatic residues such as phenylalanine, tyrosine, histidine, and tryptophan.

In Publication 2, a systematic benchmarking study was conducted to identify an optimal QM method for the quantitative description of these interactions. Through extensive comparisons against CCSD(T)/CBS reference data, MP2/TZVPP emerged as the most balanced approach, combining near-reference accuracy (RMSD ≈ 1 kJ/mol) with reasonable computational cost. This finding provided the methodological basis for generating the large and consistent data sets required for data-driven modeling.

The subsequent work (Publication 3) integrated these high-quality QM data into neural network models capable of predicting halogen···π interaction energies with near-quantum accuracy. By training on geometrically diverse halobenzene-benzene complexes, the developed model reproduced MP2-level interaction energies with excellent agreement ($R^2 \approx 0.998$) while reducing computation time by approximately eight orders of magnitude (10^8). Tests on both random-generated and PDB-derived

geometries confirmed robust generalization, and error analyses revealed a well-defined applicability domain limited to σ-hole-driven interactions. This established a new, efficient pathway for integrating physically meaningful halogen bonding representations into large-scale molecular simulations and structure-based design workflows.

Finally, Publication 4 extended the QM-AI approach completing the set of halogen···π acceptors by incorporating interactions addressing phenol, imidazole, and indole, the aromatic side chain residues of tyrosine, histidine, and tryptophan. Despite the increased chemical complexity, the neural networks maintained near-MP2 accuracy across all systems ($R^2 \approx 0.99$) and successfully transferred it to real-world PDB geometries. The retraining and extension of the models further demonstrated their scalability and adaptability to new chemical environments. This modular approach provides a generalizable framework that can be incrementally expanded to include additional interaction types and environments.

Beyond their specific focus on halogen bonding, the methods and insights developed here contribute to the broader field of noncovalent interaction modeling. The demonstrated QM-AI workflow provides a template for other interactions, such as chalcogen or pnictogen bonds, where similar challenges of accuracy and scalability arise. Moreover, integration of these models into molecular docking or free-energy frameworks could significantly improve the physical realism of computational screening and scoring functions.

In summary, this thesis establishes a conceptual and computational bridge between detailed QM understanding and data-driven modeling of σ-hole interactions. Future work will focus on expanding the chemical and environmental scope of these models, exploring cooperative effects in multi-interaction networks, and embedding them into practical workflows for virtual screening and molecular docking.

References

(1) Bissantz, C.; Kuhn, B.; Stahl, M. A Medicinal Chemist's Guide to Molecular Interactions. *J. Med. Chem.* **2010**, *53* (14), 5061-5084. DOI: 10.1021/jm100112j.

(2) Auffinger, P.; Hays, F. A.; Westhof, E.; Ho, P. S. Halogen bonds in biological molecules. *Proc. Natl. Acad. Sci. U. S. A.* **2004**, *101* (48), 16789-16794. DOI: 10.1073/pnas.0407607101.

(3) Cavallo, G.; Metrangolo, P.; Milani, R.; Pilati, T.; Priimagi, A.; Resnati, G.; Terraneo, G. The Halogen Bond. *Chem. Rev.* **2016**, *116* (4), 2478-2601. DOI: 10.1021/acs.chemrev.5b00484.

(4) Erdélyi, M. Halogen bonding in solution. *Chem. Soc. Rev.* **2012**, *41* (9), 3547-3557. DOI: 10.1039/C2CS15292D.

(5) Politzer, P.; Murray, J. S.; Clark, T. Halogen bonding and other σ-hole interactions: a perspective. *Phys. Chem. Chem. Phys.* **2013**, *15* (27), 11178-11189. DOI: 10.1039/C3CP00054K.

(6) Voth, A. R.; Shing Ho, P. The Role of Halogen Bonding in Inhibitor Recognition and Binding by Protein Kinases. *Curr. Top. Med. Chem.* **2007**, *7* (14), 1336-1348. DOI: 10.2174/156802607781696846.

(7) Shinada, N. K.; de Brevern, A. G.; Schmidtke, P. Halogens in Protein–Ligand Binding Mechanism: A Structural Perspective. *J. Med. Chem.* **2019**, *62* (21), 9341-9356. DOI: 10.1021/acs.jmedchem.8b01453.

(8) Parisini, E.; Metrangolo, P.; Pilati, T.; Resnati, G.; Terraneo, G. Halogen bonding in halocarbon–protein complexes: a structural survey. *Chem. Soc. Rev.* **2011**, *40* (5), 2267-2278. DOI: 10.1039/C0CS00177E.

(9) Clark, T.; Hennemann, M.; Murray, J. S.; Politzer, P. Halogen bonding: the σ-hole. *J. Mol. Model.* **2007**, *13* (2), 291-296. DOI: 10.1007/s00894-006-0130-2.

(10) Desiraju, G. R.; Ho, P. S.; Kloo, L.; Legon, A. C.; Marquardt, R.; Metrangolo, P.; Politzer, P.; Resnati, G.; Rissanen, K. Definition of the halogen bond (IUPAC Recommendations 2013). *Pure Appl. Chem.* **2013**, *85* (8), 1711-1713. DOI: 10.1351/PAC-REC-12-05-10.

(11) Sedlak, R.; Kolář, M. H.; Hobza, P. Polar Flattening and the Strength of Halogen Bonding. *J. Chem. Theory Comput.* **2015**, *11* (10), 4727-4732. DOI: 10.1021/acs.jctc.5b00687.

(12) Politzer, P.; Murray, J. S.; Clark, T. Halogen bonding: an electrostatically-driven highly directional noncovalent interaction. *Phys. Chem. Chem. Phys.* **2010**, *12* (28), 7748-7757, 10.1039/C004189K. DOI: 10.1039/C004189K.

(13) Wang, C.; Danovich, D.; Mo, Y.; Shaik, S. On The Nature of the Halogen Bond. *J. Chem. Theory Comput.* **2014**, *10* (9), 3726-3737. DOI: 10.1021/ct500422t.

(14) Kozuch, S.; Martin, J. M. L. Halogen Bonds: Benchmarks and Theoretical Analysis. *J. Chem. Theory Comput.* **2013**, *9* (4), 1918-1931. DOI: 10.1021/ct301064t.

(15) Riley, K. E.; Ford, C. L.; Demouchet, K. Comparison of hydrogen bonds, halogen bonds, CH···π interactions, and CX···π interactions using high-level ab initio methods. *Chem. Phys. Lett.* **2015**, *621*, 165-170. DOI: doi.org/10.1016/j.cplett.2014.12.040.

(16) Wolters, L. P.; Schyman, P.; Pavan, M. J.; Jorgensen, W. L.; Bickelhaupt, F. M.; Kozuch, S. The many faces of halogen bonding: a review of theoretical models and methods. *WIREs Comput. Mol. Sci.* **2014**, *4* (6), 523-540. DOI: 10.1002/wcms.1189.

(17) Ho, P. S. Halogen Bonding in Medicinal Chemistry: From Observation to Prediction. *Future Med. Chem.* **2017**, *9* (7), 637-640. DOI: 10.4155/fmc-2017-0052.

(18) Ford, M. C.; Ho, P. S. Computational Tools To Model Halogen Bonds in Medicinal Chemistry. *J. Med. Chem.* **2016**, *59* (5), 1655-1670. DOI: 10.1021/acs.jmedchem.5b00997.

(19) Sakai, T.; Torii, H. Substituent Effect and Its Halogen-Atom Dependence of Halogen Bonding Viewed through Electron Density Changes. *Chem. Asian J.* **2023**, *18* (3), e202201196. DOI: 10.1002/asia.202201196.

(20) Lange, A.; Heidrich, J.; Zimmermann, M. O.; Exner, T. E.; Boeckler, F. M. Scaffold Effects on Halogen Bonding Strength. *J. Chem. Inf. Model.* **2019**, *59* (2), 885-894. DOI: 10.1021/acs.jcim.8b00621.

(21) Bhattarai, S.; Sutradhar, D.; Chandra, A. K. Tuning of halogen-bond strength: Comparative role of basicity and strength of σ-hole. *J. Mol. Struct.* **2021**, *1223*, 129239. DOI: doi.org/10.1016/j.molstruc.2020.129239.

(22) Donald, K. J.; Pham, N.; Ravichandran, P. Sigma Hole Potentials as Tools: Quantifying and Partitioning Substituent Effects. *The Journal of Physical Chemistry A* **2023**, *127* (48), 10147-10158. DOI: 10.1021/acs.jpca.3c05797.

(23) Esrafili, M. D.; Mahdavinia, G.; Javaheri, M.; Sobhi, H. R. A theoretical study of substitution effects on halogen–π interactions. *Mol. Phys.* **2014**, *112* (8), 1160-1166. DOI: 10.1080/00268976.2013.837535.

(24) Wilcken, R.; Zimmermann, M. O.; Lange, A.; Zahn, S.; Boeckler, F. M. Using halogen bonds to address the protein backbone: a systematic evaluation. *J. Comput. Aided Mol. Des.* **2012**, *26* (8), 935-945. DOI: 10.1007/s10822-012-9592-8.

(25) Wilcken, R.; Zimmermann, M. O.; Lange, A.; Joerger, A. C.; Boeckler, F. M. Principles and Applications of Halogen Bonding in Medicinal Chemistry and Chemical Biology. *J. Med. Chem.* **2013**, *56* (4), 1363-1388. DOI: 10.1021/jm3012068.

(26) Rowe, R. K.; Ho, P. S. Relationships between hydrogen bonds and halogen bonds in biological systems. *Acta Crystallogr. Sect. B: Struct. Sci.* **2017**, *73* (2), 255-264. DOI: 10.1107/S2052520617003109.

(27) Udachin, K. A.; Alavi, S.; Ripmeester, J. A. Water–Halogen Interactions in Chlorine and Bromine Clathrate Hydrates: An Example of Multidirectional Halogen Bonding. *Journal of Physical Chemistry C* **2013**, *117* (27), 14176-14182. DOI: 10.1021/jp402399r.

(28) Hardegger, L. A.; Kuhn, B.; Spinnler, B.; Anselm, L.; Ecabert, R.; Stihle, M.; Gsell, B.; Thoma, R.; Diez, J.; Benz, J.; et al. Systematic Investigation of Halogen Bonding in Protein–Ligand Interactions. *Angew. Chem. Int. Ed.* **2011**, *50* (1), 314-318. DOI: 10.1002/anie.201006781.

(29) Hernandes, Z. M.; Cavalcanti, T. S. M.; Moreira, M. D. R.; de Azevedo Junior, F. W.; Leite, L. A. C. Halogen Atoms in the Modern Medicinal Chemistry: Hints for the Drug Design. *Curr. Drug Targets* **2010**, *11* (3), 303-314. DOI: 10.2174/138945010790711996.

(30) Jiang, L.; Zhang, X.; Zhou, Y.; Chen, Y.; Luo, Z.; Li, J.; Yuan, C.; Huang, M. Halogen bonding for the design of inhibitors by targeting the S1 pocket of serine proteases. *RSC Adv.* **2018**, *8* (49), 28189-28197, 10.1039/C8RA03145B. DOI: 10.1039/C8RA03145B.

(31) Metrangolo, P.; Resnati, G. *Halogen Bonding: Fundamentals and Applications*; Springer, 2008. DOI: 10.1007/978-3-540-74330-9.

(32) Metrangolo, P.; Resnati, G. *Halogen Bonding I: Impact on Materials Chemistry and Life Sciences*; Springer, 2015. DOI: 10.1007/978-3-319-14057-5.

(33) Metrangolo, P.; Resnati, G. *Halogen Bonding II: Impact on Materials Chemistry and Life Sciences*; Springer, 2015. DOI: 10.1007/978-3-319-15732-0.

(34) Parker, A. J.; Stewart, J.; Donald, K. J.; Parish, C. A. Halogen Bonding in DNA Base Pairs. *J. Am. Chem. Soc.* **2012**, *134* (11), 5165-5172. DOI: 10.1021/ja2105027.

(35) Scholfield, M. R.; Zanden, C. M. V.; Carter, M.; Ho, P. S. Halogen bonding (X-bonding): A biological perspective. *Protein Sci.* **2013**, *22* (2), 139-152. DOI: 10.1002/pro.2201.

(36) Shah, M. B.; Liu, J.; Zhang, Q.; Stout, C. D.; Halpert, J. R. Halogen−π Interactions in the Cytochrome P450 Active Site: Structural Insights into Human CYP2B6 Substrate Selectivity. *ACS Chem. Biol.* **2017**, *12* (5), 1204-1210. DOI: 10.1021/acschembio.7b00056.

(37) Sirimulla, S.; Bailey, J. B.; Vegesna, R.; Narayan, M. Halogen Interactions in Protein–Ligand Complexes: Implications of Halogen Bonding for Rational Drug Design. *J. Chem. Inf. Model.* **2013**, *53* (11), 2781-2791. DOI: 10.1021/ci400257k.

(38) Xu, Z.; Yang, Z.; Liu, Y.; Lu, Y.; Chen, K.; Zhu, W. Halogen Bond: Its Role beyond Drug–Target Binding Affinity for Drug Discovery and Development. *J. Chem. Inf. Model.* **2014**, *54* (1), 69-78. DOI: 10.1021/ci400539q.

(39) Berger, G.; Frangville, P.; Meyer, F. Halogen bonding for molecular recognition: new developments in materials and biological sciences. *Chem. Commun.* **2020**, *56* (37), 4970-4981, 10.1039/D0CC00841A. DOI: 10.1039/D0CC00841A.

(40) Vaas, S.; Zimmermann, M. O.; Schollmeyer, D.; Stahlecker, J.; Engelhardt, M. U.; Rheinganz, J.; Drotleff, B.; Olfert, M.; Lämmerhofer, M.; Kramer, M.; et al. Principles and Applications of CF2X Moieties as Unconventional Halogen Bond Donors in Medicinal Chemistry, Chemical Biology, and Drug Discovery. *J. Med. Chem.* **2023**, *66* (15), 10202-10225. DOI: 10.1021/acs.jmedchem.3c00634.

(41) Walker, M. G.; Mendez, C. G.; Ho, A. N.; Czarny, R. S.; Rappé, A. K.; Ho, P. S. Design of a halogen bond catalyzed DNA endonuclease. *Proceedings of the National Academy of Sciences* **2025**, *122* (14), e2500099122. DOI: doi:10.1073/pnas.2500099122.

(42) Wilcken, R.; Zimmermann, M. O.; Bauer, M. R.; Rutherford, T. J.; Fersht, A. R.; Joerger, A. C.; Boeckler, F. M. Experimental and Theoretical Evaluation of the Ethynyl Moiety as a Halogen Bioisostere. *ACS Chem. Biol.* **2015**, *10* (12), 2725-2732. DOI: 10.1021/acschembio.5b00515.

(43) *PDB - Small Molecule Statistics: Molecular Weight Distribution*. https://www.rcsb.org/stats/chemcomp/distribution-chem-comp-molecular-weight (accessed 2023 -).

(44) Zimmermann, M. O.; Boeckler, F. M. Targeting the protein backbone with aryl halides: systematic comparison of halogen bonding and π···π interactions using N-methylacetamide. *MedChemComm* **2016**, *7* (3), 500-505. DOI: 10.1039/C5MD00499C.

(45) Zimmermann, M. O.; Lange, A.; Boeckler, F. M. Evaluating the Potential of Halogen Bonding in Molecular Design: Automated Scaffold Decoration Using the New Scoring Function XBScore. *J. Chem. Inf. Model.* **2015**, *55* (3), 687-699. DOI: 10.1021/ci5007118.

(46) Wilcken, R.; Zimmermann, M. O.; Lange, A.; Zahn, S.; Kirchner, B.; Boeckler, F. M. Addressing Methionine in Molecular Design through Directed Sulfur–Halogen Bonds. *J. Chem. Theory Comput.* **2011**, *7* (7), 2307-2315. DOI: 10.1021/ct200245e.

(47) Koebel, M. R.; Cooper, A.; Schmadeke, G.; Jeon, S.; Narayan, M.; Sirimulla, S. S···O and S···N Sulfur Bonding Interactions in Protein–Ligand Complexes: Empirical Considerations and Scoring Function. *J. Chem. Inf. Model.* **2016**, *56* (12), 2298-2309. DOI: 10.1021/acs.jcim.6b00236.

(48) Murray, J. S.; Lane, P.; Politzer, P. Simultaneous σ-hole and hydrogen bonding by sulfur- and selenium-containing heterocycles. *Int. J. Quantum Chem.* **2008**, *108* (15), 2770-2781. DOI: 10.1002/qua.21753 (acccessed 2025/10/29).

(49) Lange, A.; Zimmermann, M. O.; Wilcken, R.; Zahn, S.; Boeckler, F. M. Targeting Histidine Side Chains in Molecular Design through Nitrogen–Halogen Bonds. *J. Chem. Inf. Model.* **2013**, *53* (12), 3178-3189. DOI: 10.1021/ci4004305.

(50) Zimmermann, M. O.; Lange, A.; Zahn, S.; Exner, T. E.; Boeckler, F. M. Using Surface Scans for the Evaluation of Halogen Bonds toward the Side Chains of Aspartate, Asparagine, Glutamate, and Glutamine. *J. Chem. Inf. Model.* **2016**, *56* (7), 1373-1383. DOI: 10.1021/acs.jcim.6b00075.

(51) Zhou, P.; Lv, J.; Zou, J.; Tian, F.; Shang, Z. Halogen–water–hydrogen bridges in biomolecules. *J. Struct. Biol.* **2010**, *169* (2), 172-182. DOI: 10.1016/j.jsb.2009.10.006.

(52) Bodnarchuk, M. S.; Viner, R.; Michel, J.; Essex, J. W. Strategies to Calculate Water Binding Free Energies in Protein–Ligand Complexes. *J. Chem. Inf. Model.* **2014**, *54* (6), 1623-1633. DOI: 10.1021/ci400674k.

(53) Karas, L. J.; Wu, C.-H.; Das, R.; Wu, J. I. C. Hydrogen bond design principles. *WIREs Comput. Mol. Sci.* **2020**, *10* (6), e1477. DOI: 10.1002/wcms.1477.

(54) Maurer, M.; Oostenbrink, C. Water in protein hydration and ligand recognition. *J. Mol. Recognit.* **2019**, *32* (12), e2810. DOI: 10.1002/jmr.2810.

(55) Ladbury, J. E. Just add water! The effect of water on the specificity of protein-ligand binding sites and its potential application to drug design. *Chem. Biol.* **1996**, *3* (12), 973-980. DOI: 10.1016/S1074-5521(96)90164-7.

(56) Poornima, C. S.; Dean, P. M. Hydration in drug design. 3. Conserved water molecules at the ligand-binding sites of homologous proteins. *J. Comput. Aided Mol. Des.* **1995**, *9* (6), 521-531. DOI: 10.1007/BF00124323.

(57) Poornima, C. S.; Dean, P. M. Hydration in drug design. 1. Multiple hydrogen-bonding features of water molecules in mediating protein-ligand interactions. *J. Comput. Aided Mol. Des.* **1995**, *9* (6), 500-512. DOI: 10.1007/BF00124321.

(58) Poornima, C. S.; Dean, P. M. Hydration in drug design. 2. Influence of local site surface shape on water binding. *J. Comput. Aided Mol. Des.* **1995**, *9* (6), 513-520. DOI: 10.1007/BF00124322.

(59) Barillari, C.; Taylor, J.; Viner, R.; Essex, J. W. Classification of Water Molecules in Protein Binding Sites. *J. Am. Chem. Soc.* **2007**, *129* (9), 2577-2587. DOI: 10.1021/ja066980q.

(60) Spyrakis, F.; Ahmed, M. H.; Bayden, A. S.; Cozzini, P.; Mozzarelli, A.; Kellogg, G. E. The Roles of Water in the Protein Matrix: A Largely Untapped Resource for Drug Discovery. *J. Med. Chem.* **2017**, *60* (16), 6781-6827. DOI: 10.1021/acs.jmedchem.7b00057.

(61) Shaltiel, S.; Cox, S.; Taylor, S. S. Conserved water molecules contribute to the extensive network of interactions at the active site of protein kinase A. *Proceedings of the National Academy of Sciences* **1998**, *95* (2), 484-491. DOI: doi:10.1073/pnas.95.2.484.

(62) Bairagya, H. R.; Mukhopadhyay, B. P.; Bhattacharya, S. Role of the conserved water molecules in the binding of inhibitor to IMPDH-II (human): A study on the water mimic inhibitor design. *J. Mol. Struct.* **2009**, *908* (1), 31-39. DOI: 10.1016/j.theochem.2009.04.037.

(63) Abel, R.; Young, T.; Farid, R.; Berne, B. J.; Friesner, R. A. Role of the Active-Site Solvent in the Thermodynamics of Factor Xa Ligand Binding. *J. Am. Chem. Soc.* **2008**, *130* (9), 2817-2831. DOI: 10.1021/ja0771033.

(64) Clarke, C.; Woods, R. J.; Gluska, J.; Cooper, A.; Nutley, M. A.; Boons, G.-J. Involvement of Water in Carbohydrate−Protein Binding. *J. Am. Chem. Soc.* **2001**, *123* (49), 12238-12247. DOI: 10.1021/ja004315q.

(65) Geschwindner, S.; Ulander, J. The current impact of water thermodynamics for small-molecule drug discovery. *Expert Opin. Drug Dis.* **2019**, *14* (12), 1221-1225. DOI: 10.1080/17460441.2019.1664468.

(66) Matricon, P.; Suresh, R. R.; Gao, Z.-G.; Panel, N.; Jacobson, K. A.; Carlsson, J. Ligand design by targeting a binding site water. *Chem. Sci.* **2021**, *12* (3), 960-968. DOI: 10.1039/D0SC04938G.

(67) Pastor, M.; Cruciani, G.; Watson, K. A. A Strategy for the Incorporation of Water Molecules Present in a Ligand Binding Site into a Three-Dimensional Quantitative Structure−Activity Relationship Analysis. *J. Med. Chem.* **1997**, *40* (25), 4089-4102. DOI: 10.1021/jm970273d.

(68) Samways, M. L.; Bruce Macdonald, H. E.; Taylor, R. D.; Essex, J. W. Water Networks in Complexes between Proteins and FDA-Approved Drugs. *J. Chem. Inf. Model.* **2023**, *63* (1), 387-396. DOI: 10.1021/acs.jcim.2c01225.

(69) Verteramo, M. L.; Ignjatović, M. M.; Kumar, R.; Wernersson, S.; Ekberg, V.; Wallerstein, J.; Carlström, G.; Chadimová, V.; Leffler, H.; Zetterberg, F.; et al. Interplay of halogen bonding and solvation in protein–ligand binding. *iScience* **2024**, *27* (4), 109636. DOI: 10.1016/j.isci.2024.109636.

(70) Lu, Y.; Liu, Y.; Li, H.; Zhu, X.; Liu, H.; Zhu, W. Energetic Effects between Halogen Bonds and Anion-π or Lone Pair-π Interactions: A Theoretical Study. *J. Phys. Chem. A* **2012**, *116* (10), 2591-2597. DOI: 10.1021/jp212522k.

(71) Matter, H.; Nazaré, M.; Güssregen, S.; Will, D. W.; Schreuder, H.; Bauer, A.; Urmann, M.; Ritter, K.; Wagner, M.; Wehner, V. Evidence for C-Cl/C-Br···π Interactions

as an Important Contribution to Protein–Ligand Binding Affinity. *Angew. Chem. Int. Ed.* **2009**, *48* (16), 2911-2916. DOI: 10.1002/anie.200806219.

(72) Mitra, D.; Bankoti, N.; Michael, D.; Sekar, K.; Row, T. N. G. C-halogen…pi interactions in nucleic acids: a database study. *Journal of Chemical Sciences* **2020**, *132* (1), 93. DOI: 10.1007/s12039-020-01794-1.

(73) Portela, S.; Fernández, I. Nature of C−I···π Halogen Bonding and its Role in Organocatalysis. *European Journal of Organic Chemistry* **2021**, *2021* (45), 6102-6110. DOI: 10.1002/ejoc.202101244 (acccessed 2025/02/06).

(74) Schollmeyer, D.; Shishkin, O. V.; Rühl, T.; Vysotsky, M. O. OH–π and halogen–π interactions as driving forces in the crystal organisations of tri-bromo and tri-iodo trityl alcohols. *CrystEngComm* **2008**, *10* (6), 715-723, 10.1039/B716442D. DOI: 10.1039/B716442D.

(75) Youn, I. S.; Kim, D. Y.; Cho, W. J.; Madridejos, J. M. L.; Lee, H. M.; Kołaski, M.; Lee, J.; Baig, C.; Shin, S. K.; Filatov, M.; et al. Halogen−π Interactions between Benzene and X2/CX4 (X = Cl, Br): Assessment of Various Density Functionals with Respect to CCSD(T). *J. Phys. Chem. A* **2016**, *120* (46), 9305-9314. DOI: 10.1021/acs.jpca.6b09395.

(76) Forni, A.; Pieraccini, S.; Rendine, S.; Gabas, F.; Sironi, M. Halogen-Bonding Interactions with π Systems: CCSD(T), MP2, and DFT Calculations. *Chemphyschem* **2012**, *13* (18), 4224-4234. DOI: 10.1002/cphc.201200605 (acccessed 2025/02/06).

(77) Řezáč, J.; Riley, K. E.; Hobza, P. Benchmark Calculations of Noncovalent Interactions of Halogenated Molecules. *J. Chem. Theory Comput.* **2012**, *8* (11), 4285-4292. DOI: 10.1021/ct300647k.

(78) Zhu, Z.; Xu, Z.; Zhu, W. Interaction Nature and Computational Methods for Halogen Bonding: A Perspective. *J. Chem. Inf. Model.* **2020**, *60* (6), 2683-2696. DOI: 10.1021/acs.jcim.0c00032.

(79) Wallnoefer, H. G.; Fox, T.; Liedl, K. R.; Tautermann, C. S. Dispersion dominated halogen–π interactions: energies and locations of minima. *Phys. Chem. Chem. Phys.* **2010**, *12* (45), 14941-14949, 10.1039/C0CP00607F. DOI: 10.1039/C0CP00607F.

(80) Yan, X. C.; Schyman, P.; Jorgensen, W. L. Cooperative Effects and Optimal Halogen Bonding Motifs for Self-Assembling Systems. *J. Phys. Chem. A* **2014**, *118* (15), 2820-2826. DOI: 10.1021/jp501553j.

(81) Groenhof, G. Introduction to QM/MM Simulations. In *Biomolecular Simulations: Methods and Protocols*, Monticelli, L., Salonen, E. Eds.; Humana Press, 2013; pp 43-66.

(82) Young, D. C. *Computational Chemistry: A Practical Guide for Applying Techniques to Real World Problems*; Wiley, 2001. DOI: 10.1002/0471220655.

(83) Jensen, F. *Introduction to Computational Chemistry*; Wiley, 2017.

(84) Mayer, I. The Hartree-Fock Method. In *Simple Theorems, Proofs, and Derivations in Quantum Chemistry*, Mayer, I. Ed.; Springer US, 2003; pp 165-225.

(85) Magnasco, V. Chapter 16 - Post-Hartree–Fock methods. In *Elementary Molecular Quantum Mechanics (Second Edition)*, Magnasco, V. Ed.; Elsevier, 2013; pp 681-722.

(86) Bartlett, R. J.; Stanton, J. F. Applications of Post-Hartree—Fock Methods: A Tutorial. In *Reviews in Computational Chemistry*, Reviews in Computational Chemistry, 1994; pp 65-169.

(87) David Sherrill, C.; Schaefer, H. F. The Configuration Interaction Method: Advances in Highly Correlated Approaches. In *Advances in Quantum Chemistry*, Löwdin, P.-O., Sabin, J. R., Zerner, M. C., Brändas, E. Eds.; Vol. 34; Academic Press, 1999; pp 143-269.

(88) Møller, C.; Plesset, M. S. Note on an Approximation Treatment for Many-Electron Systems. *Phys. Rev.* **1934**, *46* (7), 618-622. DOI: 10.1103/PhysRev.46.618.

(89) Head-Gordon, M.; Pople, J. A.; Frisch, M. J. MP2 energy evaluation by direct methods. *Chem. Phys. Lett.* **1988**, *153* (6), 503-506. DOI: 10.1016/0009-2614(88)85250-3.

(90) Pople, J. A.; Binkley, J. S.; Seeger, R. Theoretical models incorporating electron correlation. *Int. J. Quantum Chem.* **1976**, *10* (S10), 1-19. DOI: 10.1002/qua.560100802 (acccessed 2025/02/06).

(91) Pople, J. A.; Seeger, R.; Krishnan, R. Variational configuration interaction methods and comparison with perturbation theory. *Int. J. Quantum Chem.* **1977**, *12* (S11), 149-163. DOI: 10.1002/qua.560120820 (acccessed 2025/02/06).

(92) Riley, K. E.; Řezáč, J.; Hobza, P. MP2.X: a generalized MP2.5 method that produces improved binding energies with smaller basis sets. *Phys. Chem. Chem. Phys.* **2011**, *13* (47), 21121-21125, 10.1039/C1CP22525A. DOI: 10.1039/C1CP22525A.

(93) Grimme, S. Improved second-order Møller–Plesset perturbation theory by separate scaling of parallel- and antiparallel-spin pair correlation energies. *J. Chem. Phys.* **2003**, *118* (20), 9095-9102. DOI: 10.1063/1.1569242 (acccessed 2/6/2025).

(94) Grimme, S.; Goerigk, L.; Fink, R. F. Spin-component-scaled electron correlation methods. *WIREs Comput. Mol. Sci.* **2012**, *2* (6), 886-906. DOI: 10.1002/wcms.1110 (acccessed 2025/02/06).

(95) Taylor, P. R. Coupled-cluster Methods in Quantum Chemistry. In *Lecture Notes in Quantum Chemistry II: European Summer School in Quantum Chemistry*, Roos, B. O. Ed.; Springer Berlin Heidelberg, 1994; pp 125-202.

(96) Dykstra, C. E.; Frenking, G.; Kim, K. S.; Scuseria, G. E. Chapter 1 - Computing technologies, theories, and algorithms. The making of 40 years and more of theoretical and computational chemistry. In *Theory and Applications of Computational Chemistry*, Dykstra, C. E., Frenking, G., Kim, K. S., Scuseria, G. E. Eds.; Elsevier, 2005; pp 1-7.

(97) Řezáč, J.; Hobza, P. Describing Noncovalent Interactions beyond the Common Approximations: How Accurate Is the “Gold Standard,” CCSD(T) at the Complete Basis Set Limit? *J. Chem. Theory Comput.* **2013**, *9* (5), 2151-2155. DOI: 10.1021/ct400057w.

(98) Kohn, W.; Becke, A. D.; Parr, R. G. Density Functional Theory of Electronic Structure. *The Journal of Physical Chemistry* **1996**, *100* (31), 12974-12980. DOI: 10.1021/jp960669l.

(99) Nomura, Y.; Akashi, R. Density functional theory. In *Encyclopedia of Condensed Matter Physics (Second Edition)*, Chakraborty, T. Ed.; Academic Press, 2024; pp 867-878.

(100) Kohn, W.; Sham, L. J. Self-Consistent Equations Including Exchange and Correlation Effects. *Phys. Rev.* **1965**, *140* (4A), A1133-A1138. DOI: 10.1103/PhysRev.140.A1133.

(101) Staroverov, V. N.; Scuseria, G. E.; Tao, J.; Perdew, J. P. Comparative assessment of a new nonempirical density functional: Molecules and hydrogen-bonded complexes. *J. Chem. Phys.* **2003**, *119* (23), 12129-12137. DOI: 10.1063/1.1626543.

(102) Perdew, J. P.; Chevary, J. A.; Vosko, S. H.; Jackson, K. A.; Pederson, M. R.; Singh, D. J.; Fiolhais, C. Atoms, molecules, solids, and surfaces: Applications of the generalized gradient approximation for exchange and correlation. *Phys. Rev. B* **1992**, *46* (11), 6671-6687. DOI: 10.1103/PhysRevB.46.6671.

(103) Becke, A. D. Density-functional thermochemistry. III. The role of exact exchange. *J. Chem. Phys.* **1993**, *98* (7), 5648-5652. DOI: 10.1063/1.464913 (acccessed 2/6/2025).

(104) Lee, C.; Yang, W.; Parr, R. G. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Phys. Rev. B* **1988**, *37* (2), 785-789. DOI: 10.1103/PhysRevB.37.785.

(105) Zhao, Y.; Truhlar, D. G. The M06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: two new functionals and systematic testing of four M06-class functionals and 12 other functionals. *Theor. Chem. Acc.* **2008**, *120* (1), 215-241. DOI: 10.1007/s00214-007-0310-x.

(106) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *J. Chem. Phys.* **2010**, *132* (15), 154104. DOI: 10.1063/1.3382344.

(107) Helgaker, T.; Jorgensen, P.; Olsen, J. *Molecular Electronic-Structure Theory*; Wiley, 2014.

(108) Schäfer, A.; Huber, C.; Ahlrichs, R. Fully optimized contracted Gaussian basis sets of triple zeta valence quality for atoms Li to Kr. *J. Chem. Phys.* **1994**, *100* (8), 5829-5835. DOI: 10.1063/1.467146.

(109) Hellweg, A.; Rappoport, D. Development of new auxiliary basis functions of the Karlsruhe segmented contracted basis sets including diffuse basis functions (def2-SVPD, def2-TZVPPD, and def2-QVPPD) for RI-MP2 and RI-CC calculations. *Phys. Chem. Chem. Phys.* **2015**, *17* (2), 1010-1017, 10.1039/C4CP04286G. DOI: 10.1039/C4CP04286G.

(110) Dunning, T. H., Jr. Gaussian basis sets for use in correlated molecular calculations. I. The atoms boron through neon and hydrogen. *J. Chem. Phys.* **1989**, *90* (2), 1007-1023. DOI: 10.1063/1.456153 (acccessed 2/6/2025).

(111) Woon, D. E.; Dunning, T. H., Jr. Gaussian basis sets for use in correlated molecular calculations. IV. Calculation of static electrical response properties. *J. Chem. Phys.* **1994**, *100* (4), 2975-2988. DOI: 10.1063/1.466439 (acccessed 2/6/2025).

(112) Halkier, A.; Helgaker, T.; Jørgensen, P.; Klopper, W.; Koch, H.; Olsen, J.; Wilson, A. K. Basis-set convergence in correlated calculations on Ne, N2, and H2O. *Chem. Phys. Lett.* **1998**, *286* (3), 243-252. DOI: 10.1016/S0009-2614(98)00111-0.

(113) Boys, S. F.; Bernardi, F. The calculation of small molecular interactions by the differences of separate total energies. Some procedures with reduced errors. *Mol. Phys.* **1970**, *19* (4), 553-566. DOI: 10.1080/00268977000101561.

(114) Alvarez-Idaboy, J. R.; Galano, A. Counterpoise corrected interaction energies are not systematically better than uncorrected ones: comparison with CCSD(T) CBS

extrapolated values. *Theor. Chem. Acc.* **2010**, *126* (1), 75-85. DOI: 10.1007/s00214-009-0676-z.

(115) Muegge, I.; Rarey, M. Small Molecule Docking and Scoring. In *Reviews in Computational Chemistry*, 2001; pp 1-60.

(116) Ferrara, P.; Gohlke, H.; Price, D. J.; Klebe, G.; Brooks, C. L. Assessing Scoring Functions for Protein−Ligand Interactions. *J. Med. Chem.* **2004**, *47* (12), 3032-3047. DOI: 10.1021/jm030489h.

(117) Grinter, S. Z.; Zou, X. Challenges, Applications, and Recent Advances of Protein-Ligand Docking in Structure-Based Drug Design. *Molecules* **2014**, *19* (7), 10150-10176.

(118) Liu, J.; Wang, R. Classification of Current Scoring Functions. *J. Chem. Inf. Model.* **2015**, *55* (3), 475-482. DOI: 10.1021/ci500731a.

(119) Jalaie, M.; Fanfrlík, J.; Pecina, A.; Lepšík, M.; Řezáč, J. Comparative Analysis of Quantum-Mechanical and Standard Single-Structure Protein–Ligand Scoring Functions with MD-Based Free Energy Calculations. *J. Chem. Inf. Model.* **2025**, *65* (15), 8127-8136. DOI: 10.1021/acs.jcim.5c00604.

(120) Guedes, I. A.; Pereira, F. S. S.; Dardenne, L. E. Empirical Scoring Functions for Structure-Based Virtual Screening: Applications, Critical Aspects, and Challenges. *Front. Pharmacol.* **2018**, *Volume 9 - 2018*, Review. DOI: 10.3389/fphar.2018.01089.

(121) Huang, S.-Y.; Grinter, S. Z.; Zou, X. Scoring functions and their evaluation methods for protein–ligand docking: recent advances and future directions. *Phys. Chem. Chem. Phys.* **2010**, *12* (40), 12899-12908, 10.1039/C0CP00151A. DOI: 10.1039/C0CP00151A.

(122) Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J. Comput. Chem.* **1998**, *19* (14), 1639-1662. DOI: 10.1002/(SICI)1096-987X(19981115)19:14<1639::AID-JCC10>3.0.CO;2-B (acccessed 2025/10/30).

(123) Huey, R.; Morris, G. M.; Olson, A. J.; Goodsell, D. S. A semiempirical free energy force field with charge-based desolvation. *J. Comput. Chem.* **2007**, *28* (6), 1145-1152. DOI: 10.1002/jcc.20634 (acccessed 2025/10/30).

(124) Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. Development and validation of a genetic algorithm for flexible docking11Edited by F. E. Cohen. *J. Mol. Biol.* **1997**, *267* (3), 727-748. DOI: 10.1006/jmbi.1996.0897.

(125) Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. A Fast Flexible Docking Method using an Incremental Construction Algorithm. *J. Mol. Biol.* **1996**, *261* (3), 470-489. DOI: 10.1006/jmbi.1996.0477.

(126) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; et al. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *J. Med. Chem.* **2004**, *47* (7), 1739-1749. DOI: 10.1021/jm0306430.

(127) Muegge, I.; Martin, Y. C. A General and Fast Scoring Function for Protein−Ligand Interactions: A Simplified Potential Approach. *J. Med. Chem.* **1999**, *42* (5), 791-804. DOI: 10.1021/jm980536j.

(128) Muegge, I. PMF Scoring Revisited. *J. Med. Chem.* **2006**, *49* (20), 5895-5902. DOI: 10.1021/jm050038s.

(129) Gohlke, H.; Hendlich, M.; Klebe, G. Knowledge-based scoring function to predict protein-ligand interactions11Edited by R. Huber. *J. Mol. Biol.* **2000**, *295* (2), 337-356. DOI: 10.1006/jmbi.1999.3371.

(130) Cang, Z.; Mu, L.; Wei, G.-W. Representability of algebraic topology for biomolecules in machine learning based scoring and virtual screening. *PLOS Computational Biology* **2018**, *14* (1), e1005929. DOI: 10.1371/journal.pcbi.1005929.

(131) Stepniewska-Dziubinska, M. M.; Zielenkiewicz, P.; Siedlecki, P. Development and evaluation of a deep learning model for protein–ligand binding affinity prediction. *Bioinformatics* **2018**, *34* (21), 3666-3674. DOI: 10.1093/bioinformatics/bty374 (acccessed 10/30/2025).

(132) Boyles, F.; Deane, C. M.; Morris, G. M. Learning from the ligand: using ligand-based features to improve binding affinity prediction. *Bioinformatics* **2019**, *36* (3), 758-764. DOI: 10.1093/bioinformatics/btz665 (acccessed 10/30/2025).

(133) Li, J.; Fu, A.; Zhang, L. An Overview of Scoring Functions Used for Protein–Ligand Interactions in Molecular Docking. *Interdisciplinary Sciences: Computational Life Sciences* **2019**, *11* (2), 320-328. DOI: 10.1007/s12539-019-00327-w.

(134) Sánchez-Cruz, N.; Medina-Franco, J. L.; Mestres, J.; Barril, X. Extended connectivity interaction features: improving binding affinity prediction through chemical description. *Bioinformatics* **2020**, *37* (10), 1376-1382. DOI: 10.1093/bioinformatics/btaa982 (acccessed 10/30/2025).

(135) Xiong, G.-L.; Ye, W.-L.; Shen, C.; Lu, A.-P.; Hou, T.-J.; Cao, D.-S. Improving structure-based virtual screening performance via learning from scoring function components. *Brief. Bioinform.* **2020**, *22* (3). DOI: 10.1093/bib/bbaa094 (acccessed 10/30/2025).

(136) Dhakal, A.; McKay, C.; Tanner, J. J.; Cheng, J. Artificial intelligence in the prediction of protein–ligand interactions: recent advances and future directions. *Brief. Bioinform.* **2021**, *23* (1). DOI: 10.1093/bib/bbab476 (acccessed 10/30/2025).

(137) Meli, R.; Morris, G. M.; Biggin, P. C. Scoring Functions for Protein-Ligand Binding Affinity Prediction Using Structure-based Deep Learning: A Review. *Frontiers in Bioinformatics* **2022**, *Volume 2 - 2022*, Review. DOI: 10.3389/fbinf.2022.885983.

(138) Zhang, Z.; Chen, L.; Zhong, F.; Wang, D.; Jiang, J.; Zhang, S.; Jiang, H.; Zheng, M.; Li, X. Graph neural network approaches for drug-target interactions. *Curr. Opin. Struct. Biol.* **2022**, *73*, 102327. DOI: 10.1016/j.sbi.2021.102327.

(139) Zhao, L.; Zhu, Y.; Wang, J.; Wen, N.; Wang, C.; Cheng, L. A brief review of protein–ligand interaction prediction. *Computational and Structural Biotechnology Journal* **2022**, *20*, 2831-2838. DOI: 10.1016/j.csbj.2022.06.004.

(140) Jiménez, J.; Škalič, M.; Martínez-Rosell, G.; De Fabritiis, G. KDEEP: Protein–Ligand Absolute Binding Affinity Prediction via 3D-Convolutional Neural Networks. *J. Chem. Inf. Model.* **2018**, *58* (2), 287-296. DOI: 10.1021/acs.jcim.7b00650.

(141) Zheng, L.; Fan, J.; Mu, Y. OnionNet: a Multiple-Layer Intermolecular-Contact-Based Convolutional Neural Network for Protein–Ligand Binding Affinity Prediction. *ACS Omega* **2019**, *4* (14), 15956-15965. DOI: 10.1021/acsomega.9b01997.

(142) Wang, Z.; Zheng, L.; Liu, Y.; Qu, Y.; Li, Y.-Q.; Zhao, M.; Mu , Y.; Li , W. OnionNet-2: A Convolutional Neural Network Model for Predicting Protein-Ligand Binding Affinity Based on Residue-Atom Contacting Shells. *Frontiers in Chemistry* **2021**, *Volume 9 - 2021*, Original Research. DOI: 10.3389/fchem.2021.753002.

(143) Biau, G. Analysis of a random forests model. *J. Mach. Learn. Res.* **2012**, *13* (null), 1063–1095.

(144) Kern, C.; Klausch, T.; Kreuter, F. Tree-based Machine Learning Methods for Survey Research. *Surv Res Methods* **2019**, *13* (1), 73-93. From NLM.

(145) Cristianini, N.; Shawe-Taylor, J. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*; Cambridge University Press, 2000. DOI: DOI: 10.1017/CBO9780511801389.

(146) Vovk, V. Kernel Ridge Regression. In *Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik*, Schölkopf, B., Luo, Z., Vovk, V. Eds.; Springer Berlin Heidelberg, 2013; pp 105-116.

(147) Zou, J.; Han, Y.; So, S. S. Overview of artificial neural networks. *Methods Mol. Biol.* **2008**, *458*, 15-23. DOI: 10.1007/978-1-60327-101-1_2 From NLM.

(148) Zhang, Z. A gentle introduction to artificial neural networks. *Ann Transl Med* **2016**, *4* (19), 370. DOI: 10.21037/atm.2016.06.20 From NLM.

(149) Han, S. H.; Kim, K. W.; Kim, S.; Youn, Y. C. Artificial Neural Network: Understanding the Basic Concepts without Mathematics. *Dement Neurocogn Disord* **2018**, *17* (3), 83-89. DOI: 10.12779/dnd.2018.17.3.83 From NLM.

(150) Dubey, S. R.; Singh, S. K.; Chaudhuri, B. B. Activation functions in deep learning: A comprehensive survey and benchmark. *Neurocomputing* **2022**, *503*, 92-108. DOI: 10.1016/j.neucom.2022.06.111.

(151) Xu, B.; Wang, N.; Chen, T.; Li, M. Empirical evaluation of rectified activations in convolutional network. *arXiv preprint arXiv:1505.00853* **2015**.

(152) Kingma, D. P. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* **2014**.

(153) Reyad, M.; Sarhan, A. M.; Arafa, M. A modified Adam algorithm for deep neural network optimization. *Neural Computing and Applications* **2023**, *35* (23), 17095-17112. DOI: 10.1007/s00521-023-08568-z.

(154) Ruder, S. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747* **2016**.

(155) Umeno, Y.; Kubo, A. Prediction of electronic structure in atomistic model using artificial neural network. *Computational Materials Science* **2019**, *168*, 164-171. DOI: doi.org/10.1016/j.commatsci.2019.06.005.

(156) Zhang, Y.; Li, S.; Meng, K.; Sun, S. Machine Learning for Sequence and Structure-Based Protein–Ligand Interaction Prediction. *J. Chem. Inf. Model.* **2024**, *64* (5), 1456-1472. DOI: 10.1021/acs.jcim.3c01841.

(157) Unke, O. T.; Meuwly, M. PhysNet: A Neural Network for Predicting Energies, Forces, Dipole Moments, and Partial Charges. *J. Chem. Theory Comput.* **2019**, *15* (6), 3678-3693. DOI: 10.1021/acs.jctc.9b00181.

(158) Glick, Z. L.; Metcalf, D. P.; Glick, C. S.; Spronk, S. A.; Koutsoukas, A.; Cheney, D. L.; Sherrill, C. D. A physics-aware neural network for protein–ligand interactions with

quantum chemical accuracy. *Chem. Sci.* **2024**, *15* (33), 13313-13324, 10.1039/D4SC01029A. DOI: 10.1039/D4SC01029A.

(159) Tong, S.; Lai, F. Artificial Neural Network-Based Approach for Surface Energy Prediction. In *Recent Advances in Neuromorphic Computing*, Bai, K. J., Yi, Y. Eds.; IntechOpen, 2024.

(160) Bose, S.; Dhawan, D.; Nandi, S.; Sarkar, R. R.; Ghosh, D. Machine learning prediction of interaction energies in rigid water clusters. *Phys. Chem. Chem. Phys.* **2018**, *20* (35), 22987-22996, 10.1039/C8CP03138J. DOI: 10.1039/C8CP03138J.

(161) Shaw, R. A.; Hill, J. G. A Simple Model for Halogen Bond Interaction Energies. In *Inorganics*, 2019; Vol. 7.

(162) Samuel, H. S.; Nweke-Maraizu, U.; Etim, E. E. Machine learning for characterizing halogen bonding interactions. *Faculty of Natural and Applied Sciences Journal of Scientific Innovations* **2023**, *5* (1), 102-114. (acccessed 2025/07/11).

(163) Devore, D. P.; Shuford, K. L. Data and Molecular Fingerprint-Driven Machine Learning Approaches to Halogen Bonding. *J. Chem. Inf. Model.* **2024**. DOI: 10.1021/acs.jcim.4c01427.

(164) Yang, Y.; Fan, X.; Liu, Y.; Ye, D.; Liu, C.; Yang, H.; Su, Z.; Zhang, Y.; Liu, Y. Function and inhibition of DYRK1A: Emerging roles of treating multiple human diseases. *Biochem. Pharmacol.* **2023**, *212*, 115521. DOI: 10.1016/j.bcp.2023.115521.

(165) Meine, R.; Becker, W.; Falke, H.; Preu, L.; Loaëc, N.; Meijer, L.; Kunick, C. Indole-3-Carbonitriles as DYRK1A Inhibitors by Fragment-Based Drug Design. *Molecules* **2018**, *23* (2), 64. DOI: 10.3390/molecules23020064.

(166) PDB:8R8E. 2024. https://www.rcsb.org/structure/8R8E (accessed 2024-2025).

(167) *Schrödinger Release 2023-2: Desmond Molecular Dynamics System, D. E. Shaw Research, New York, NY, 2023. Maestro-Desmond Interoperability Tools, Schrödinger, New York, NY, 2023.*; (accessed.

(168) TURBOMOLE V7.7.1 2019, a development of University of Karlsruhe and Forschungszentrum Karlsruhe GmbH, 1989-2007, TURBOMOLE GmbH, since 2007.

Appendix A: Publication 1

Halogen Bonding on Water ─ A Drop in the Ocean?

Marc U. Engelhardt, Markus O. Zimmermann, Marcel Dammann, Jason Stahlecker, Antti Poso, Thales Kronenberger, Conrad Kunick, Thilo Stehle, Frank M. Boeckler;

Journal of Chemical Theory and Computation, 2024

DOI: 10.1021/acs.jctc.4c00834

JCTC Journal of Chemical Theory and Computation

pubs.acs.org/JCTC Article

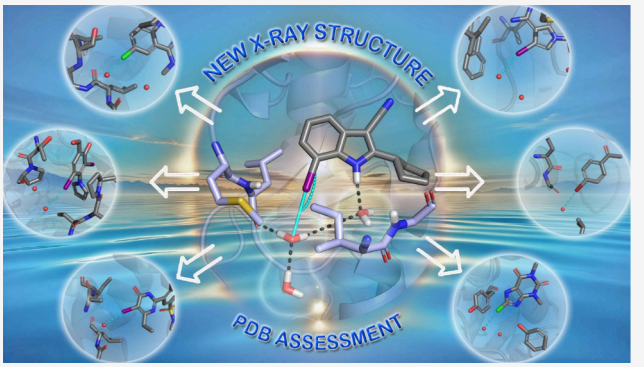
Halogen Bonding on Water—A Drop in the Ocean?

Marc U. Engelhardt, Markus O. Zimmermann,* Marcel Dammann, Jason Stahlecker, Antti Poso, Thales Kronenberger, Conrad Kunick, Thilo Stehle, and Frank M. Boeckler*

Cite This: *J. Chem. Theory Comput.* 2024, 20, 10716−10730 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Halogen bonding is a valuable interaction in drug design, offering an unconventional way to influence affinity and selectivity by leveraging the halogen atoms' ability to form directional bonds. The present study evaluates halogen−water interactions within protein binding sites, demonstrating that targeting a water molecule via halogen bonding can in specific cases contribute beneficially to ligand binding. In solving and examining the crystal structure of 2-cyclopentyl-7-iodo-1*H*-indole-3-carbonitrile bound to DYRK1a kinase, we identified a notable iodine−water interaction, where water accepts a halogen bond with good geometric and energetic features. This starting point triggered further investigations into the prevalence of such interactions across various halogen-bearing ligands (chlorine, bromine, iodine) in the PDB. Using QM calculations (MP2/TZVPP), we highlight the versatility and potential benefits of such halogen−water interactions, particularly when the water molecule is a stable part of the binding site's structured environment. While the interaction energies with water are lower compared to other typical halogen bond acceptors, we deem this different binding strength essential for reducing desolvation costs. We suggest that "interstitial" water molecules, as stable parts of the binding site engaging in multiple strong interactions, could be prime targets for halogen bonding. Further systematic studies, combining high-resolution crystal structures and quantum chemistry, are required to scrutinize whether halogen bonding on water is more than a "drop in the ocean".



INTRODUCTION

A thorough understanding of all contributions to protein−ligand recognition is of fundamental importance for modern drug discovery and design. While there is numerous research on various more classical molecular interactions (such as hydrogen bonds, van der Waals contacts, ion-pairs, ion-dipole interactions, cation···π, C−H···π, or π···π interactions), other interactions such as halogen and chalcogen bonds, both so-called σ-hole interactions, have just more recently gained attention and are not fully understood, yet.[1−7]

Halogen bonding (XB) has been compared to hydrogen bonding, because of its high directionality, the necessity to meet rather strict geometric requirements to hit the sweet spot of the interaction energy, its versatility to interact with a multitude of quite different partners in the binding site, and its ability to engage in a network of interactions (including hydrogen bonds). Hence, XBs can offer novel and useful principles and opportunities for molecular recognition and drug design.[8−13] In a nutshell, the σ-hole is an electron-deficient and, thus, positively charged region on the halogen's surface typically in elongation of the R−X axis.[5,14−16] The σ-hole facilitates attractive interactions (XBs) between this electrophilic region on the halogen atom (X), typically chlorine, bromine, or iodine, and a nucleophilic region in another molecular entity, e.g. the protein binding site. It should be noted that this strongly anisotropic electron distribution leads to a high lateral electron density, surrounding the halogen perpendicular to the R−X axis. This effect is important for the high directionality of the XB interaction. In systems, where R exerts a strongly electron-withdrawing effect on the halogen X, such strong "tuning effects" can lead to a great enhancement of the strength of the XB and the accessibility of the σ-hole, while the ability to accept hydrogen bonds laterally can vanish completely.[17,18] Consequently, also directionality can be impaired. Often a chlorine, bromine or iodine atom is reduced to its capability of donating XBs through their σ-hole with nucleophilic binding partners, because it is the more unexpected interaction mode. Still, the versatility to donate XBs and accept hydrogen bonds, as well as to integrate into an interaction network with both, can be important for its applicability.[9] To date, different nucleophilic interaction partners in the protein binding site accepting halogen bonds have been studied systematically, including backbone carbonyls

Received: June 28, 2024
Revised: September 9, 2024
Accepted: September 10, 2024
Published: September 18, 2024

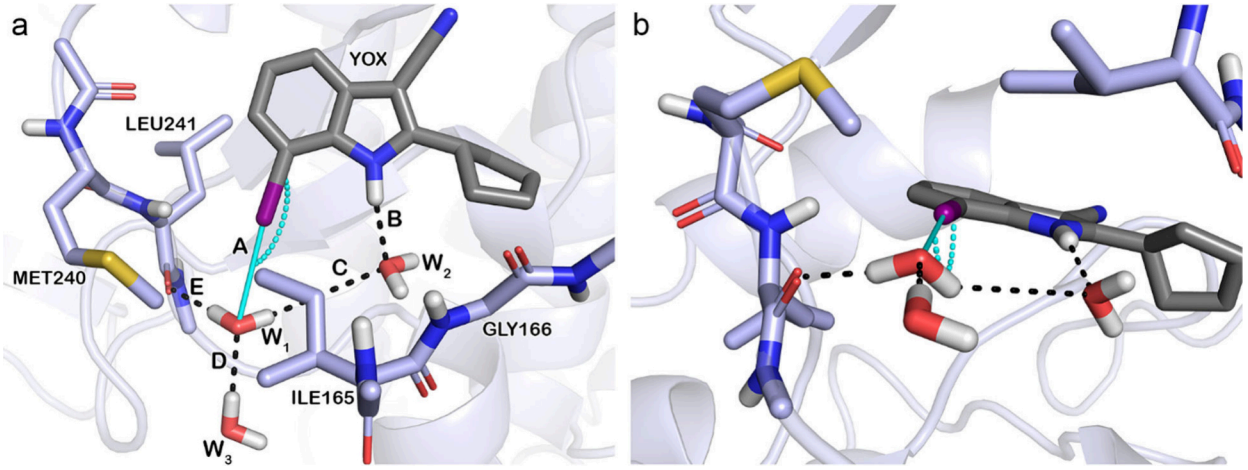


Figure 1. (a) Crystal structure of DYRK1a kinase (PDB: 8R8E) and the compound YOX. Interaction A: Compound YOX forming a favorable I···O_{water} interaction to the water residue W_1 (HOH618) with an interaction distance $d_{I\cdots O}$ = 3.35 Å and an interaction angle $\alpha_{C-X\cdots O}$ = 168.3° (cyan). Interaction B: The compound forms a hydrogen bond to another water molecule W_2 (HOH761) with $d_{H\cdots O}$ = 2.2 Å. Interaction C: Hydrogen bond between W_1 and W_2 with $d_{H\cdots O}$ = 3.9 Å. Interaction D: Hydrogen bond between W_1 and W_3 (HOH800) at $d_{H\cdots O}$ = 2.2 Å. Interaction E: Hydrogen bond between W_1 and the backbone carbonyl of LEU241 at $d_{H\cdots O}$ = 1.7 Å. (b) Different orientation of the complex. The figure was prepared using PyMOL.

and the π-surface of the peptide bond,[19,20] the sulfur atom[21] in methionine, the nitrogen atoms[22] in histidine, carboxylate[23] (aspartate/glutamate) and carboxamide moieties[23] (asparagine/glutamine). While the relevance of π-systems (aromatic side chains of tyrosine, phenylalanine, histidine, and tryptophan) and hydroxyl groups as XB acceptors has been highlighted,[13,24,25] systematic studies are still work in progress.

Similar to the hydroxyl functions in serine, threonine, or tyrosine residues, halogen bonds evidently could be formed with the oxygen atom of water molecules. In 2010, Zhou et al.[26] analyzed the interaction of halogenated nucleic acids, proteins, or small molecules with water molecules in crystal structures. Based on angle thresholds, they identified halogen−water−hydrogen (XWH) bridges, which could have an impact on protein, DNA, or RNA stability, and facilitate particular folding of these biomolecules. In these contacts, a halogen bond replaces a hydrogen bond in a water-mediated interaction. They further extend their XWH bridge concept to the recognition of small molecules in their biomolecular binding sites. Despite this pioneering work on halogen−water interactions, awareness of this topic is still limited in drug discovery. One reason might be that they have been traditionally perceived as surrogate hydrogen bond donors onto halogens, underestimating the possibility of engaging in halogen bonds. Since water is ubiquitous in the binding site of proteins, it plays a crucial role in many aspects of drug design, e.g., by mediating interactions between ligands and proteins.[27−34] Ligands are typically designed to replace energetically unfavorable water molecules. Displacing particularly ordered water molecules at hydrophobic surfaces of the protein into the bulk solvent can have a high influence on a ligand's binding affinity by increasing entropy.[35−39] On the other hand, tightly bound water molecules can be conserved as "interstitial waters" and can become especially important for the ligand's binding mode, because they are to be treated as part of the binding site.[40,41] "Trapping" a water molecule in the binding site can entropically be unfavorable but enthalpically favorable as water can form multiple interactions with its surroundings. An increase in binding affinity is thereby only achieved if the enthalpic contribution overcompensates the entropic penalty.[42]

Based on a series of indole-3-carbonitriles bearing a chlorine, bromine, or iodine atom in position 7, identified through a fragmentation approach by Meine et al.,[43] we were able to solve the crystal structure of 2-cyclopentyl-7-iodo-1*H*-indole-3-carbonitrile (**6s**,[43] Chemical ID in PDB structure 8R8E: YOX) in complex with DYRK1a. In this crystal structure, we observed a network of halogen and hydrogen bonds involving "interstitial water molecules", inspiring us to conduct a more systematic assessment of such interactions in the protein data bank (PDB).[44] A water molecule lies in the immediate vicinity of the ligand's iodine atom indicating a potential halogen bond. The interaction distance $d_{I\cdots O}$ and σ-hole angle $\alpha_{C-I\cdots O}$ between the iodine atom and the water's oxygen atom are similar to previously described favorable halogen bond complexes.[45]

Starting from a detailed analysis of our newly obtained crystal structure of Compound YOX with DYRK1a (PDB: 8R8E), we extend this analysis of halogen−water interactions to all protein crystal structures in the PDB. Using custom Python[46]/PyMOL[47] scripts, we identified potentially interesting water molecules in close proximity to halogenated ligands providing insights into their occurrence. Using quantum mechanical (QM) calculations, we optimized and analyzed the hydrogen bond networks for each complex to assess the interaction energy of the potential halogen bond. These certainly depend on various geometric parameters including orientation and hydrogen bond patterns of the individual water molecules. We highlight the frequency of combinations of distances $d_{X\cdots O}$ and σ-hole angles $\alpha_{C-X\cdots O}$ for these halogen−water contacts and their calculated adduct formation energies. Likewise, spherical distribution patterns around the oxygen of the water molecule are analyzed with respect to the type of halogen atom and their relevance for determining the nature of the interaction as halogen or hydrogen bonding is discussed.

RESULTS AND DISCUSSION

DYRK1a Crystal Structure. The Dual-specificity tyrosine phosphorylation-regulated kinase 1A (DYRK1a) is an

https://doi.org/10.1021/acs.jctc.4c00834
J. Chem. Theory Comput. 2024, 20, 10716−10730

important regulator in proliferation and differentiation of nerve cells and plays a crucial role in the development of Down syndrome and Alzheimer's disease.[48] We solved the crystal structure of DYRK1a kinase containing the inhibitor 2-cyclopentyl-7-iodo-1*H*-indole-3-carbonitrile (PDB: 8R8E).

The crystal structure provided evidence for a water molecule in close contact with the iodine atom of the ligand YOX. The interaction distance between the iodine atom and the oxygen of the water molecule and the σ-hole angle were measured as $d_{I\cdots O}$ = 3.35 Å and $\alpha_{C-I\cdots O}$ = 168.3°, respectively, suggesting a potentially favorable halogen bond. The addressed water molecule is only observed in chain A of the protein, raising doubts about its presence in that specific location across other chains. However, examination of the electron density strongly supports the existence of this water molecule. An analysis of the B-factors reveals variations in the flexibility of the binding site (Individual values can be found in Table S2 in the Supporting Information). Notably, the B-factors for the backbone carbonyl of the nearby Leucine 241, which is part of the hinge region, increase progressively across chains A to D. This indicates greater flexibility of Leucine 241, which likely influences the mobility of the water molecule and may explain its exclusive observation in chain A.

Subsequently, a section of the binding site, including the ligand, the interacting water molecule, and all surrounding amino acid and water residues within 4 Å, were extracted and protonated using the Protein Preparation Wizard of the Schrödinger Suite.[49] In this process, different hydrogen atom positions are sampled, clustered and optimized with respect to their ability to form hydrogen bonds and participate in the bond network. Finally, the most likely hydrogen bond network is retrieved as an initial guess and, subsequently, optimized at a TPSS[50]-D3[51]/TZVP[52] level of theory using TURBOMOLE.[53] Heavy atoms were kept frozen to maintain the experimentally determined binding situation. From the QM-optimized structure (Figure 1), the interaction geometries of different interaction partners were extracted. To obtain the adduct formation energy, MP2[54]/TZVPP[52] single-point calculations of the individual interaction partners were conducted. The halogen bond (Figure 1a, A) between the iodine and the oxygen atom shows a strength of −8.44 kJ/mol for the original ligand YOX. In previous systematic evaluations, we used a simple halobenzene as a model system for a not tuned haloaryl system.[19−23,45] Replacing the ligand with the model iodobenzene yields a similar interaction strength of −8.54 kJ/mol. The halogen bond acceptor W_1 forms a multitude of polar interactions with the surrounding binding site moieties. The hydrogen bond (Figure 1a, B) between the ligand and water W_2 shows a strength of −23.7 kJ/mol, while the interaction (Figure 1a, C) between the two water molecules W_1 and W_2 is only weak with a strength of −3.46 kJ/mol, due to the rather large distance of close to 4 Å. Further, the hydrogen bond (Figure 1a, D) between W_1 and W_3 exhibits a strength of −19.14 kJ/mol, while the interaction between W_1 and the nearby leucine (Figure 1a, E) shows a strength of −11.78 kJ/mol. It should be noted that the interaction energies here are provided as the simple adduction formation energies of two directly interacting molecules without taking into account the possible synergistic effects of the entire network, e.g., with respect to mutual polarization.

PDB Scan for Halogen−Water Contacts. To identify further ligand-halogen contacts with water molecules, a PDB scan was conducted. Compared to the analysis performed by Zhou et al.,[26] data availability has increased more than 4-fold over the past 15 years. In addition, we completely focused on protein−ligand recognition from a drug discovery perspective, thus, excluding all types of halogenated biomolecular building blocks. 197120 crystal structures (as of March 2023) were analyzed with 7635 (3.87%) unique structures containing ligands that bear chlorine, bromine, or iodine moieties (Table 1). Fluorine contacts were not considered, since fluorine naturally does not possess a σ-hole and therefore will not be able to engage in a halogen bond, except for extreme cases where fluorine itself is heavily tuned.[55] Halogen···water contacts were reported if they were located within 4 Å around the halogen atom. A total of 3780 water contacts in 1726 unique PDB structures were found. The difference derives either from multiple addressed water molecules within the same chain or multiple contacts in several different chains of the crystal structure.

Table 1. PDB Scan for Halogenated Ligands

Halogenated Ligands	Total	Chlorine	Bromine	Iodine
Unique PDB IDs	7635	5656	1350	454
Percentage		77.15%	18.41%	6.19%

With 74.55% of all contacts, the majority is attributed to chlorine, followed by 19.79% to bromine and 5.66% to iodine (Table 2). Figure 2 shows the distance and angle distribution of all water contacts. Distances below 2.0 Å were neglected indicating clashes or artifacts rather than favorable contacts. Chlorine (green), bromine (dark red), and iodine (purple) contacts are plotted separately (A version of the figure for readers with red-green color vision deficiency can be found in the Supporting Information, Figure S2). For better comparability between the halogens, the relative occurrence was measured. The top histogram shows the relative angle distribution in bins of 10° from 0° to 180° for chlorine, bromine, and iodine. The sum of each bin is normalized with the sum of all data points of the respective halogen. The right histogram shows relative distance distribution in bins of 0.1 Å from 2.0 Å to 4.0 Å, also for chlorine, bromine, and iodine. Bins are also normalized by the sum of all data points of the respective halogen. Most chlorine contacts are in a region of 3.5 Å to 3.9 Å with angles $\alpha_{C-X\cdots O}$ of 80° to 130° with bins of 10−12%. Bromine shows a similar distribution. For iodine, most of the contacts lie in a region of 3.5 Å to 3.7 Å with a higher relative distribution in this region (bins of 12−14%). Angle distribution also peaks at values of $\alpha_{C-X\cdots O}$ = 90° to 110° (15−16%). To be classified as a σ-hole interaction, certain interaction distance and angle criteria need to be fulfilled. Wilcken et al.[45] evaluated the interaction between halogenated ligands and the carbonyl oxygen atom of the protein backbone systematically. Optimal interaction geometries were found to have a distance $d_{X\cdots O}$ of around 3.1 Å and an interaction angle $\alpha_{C-X\cdots O}$ of 180°. Gradually deviating from the optimal halogen bond interaction distance or interaction angle led to increasing energy loss. For an interaction distance of about 4 Å, an energy loss of more than 50% is expected. For deviations of more than 40° from the optimal σ-hole angle ($\alpha_{C-X\cdots O}$ = 180°), no significant attractive interaction could be found. To focus only on relevant σ-hole interactions, we applied the following restrictions for our current analysis: the interaction distance between the halogen atom and the water's oxygen atom $d_{X\cdots O}$ must be <4 Å and the interaction angle $\alpha_{C-X\cdots O}$ must be >140°.

https://doi.org/10.1021/acs.jctc.4c00834
J. Chem. Theory Comput. 2024, 20, 10716−10730

Table 2. PDB Scan Results of Water Contacts

Water contacts	Total	Chlorine	Bromine	Iodine
Number of contacts with $d_{X\cdots O}$ < 4 Å (in unique PDB IDs)	3780 (1726)	2818 (1269)	748 (368)	214 (102)
Percentage of contacts		74.6%	19.8%	5.7%
Hits $\alpha_{C-X\cdots O}$ > 140° (% of contacts)	741 (19.6%)	561 (19.91%)	144 (19.25%)	36 (16.82%)
Hits $\alpha_{C-X\cdots O}$ > 160° (% of contacts)	246 (6.51%)	178 (6.32%)	56 (7.49%)	12 (5.61%)

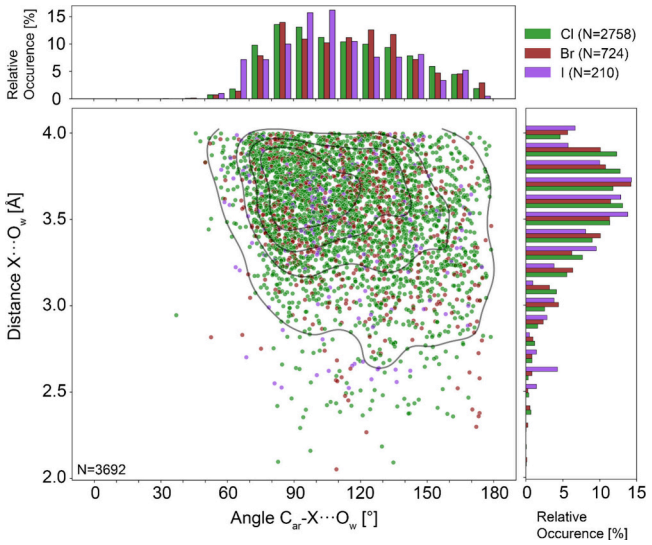


Figure 2. Distribution of halogen−water contacts in crystal structures of the PDB. Distance $d_{X\cdots O}$ is plotted against the angle $\alpha_{C-X\cdots O}$. Chlorine (green), bromine (dark red), and iodine (purple) contacts are measured independently. Contour lines show the density distribution of all data points combined. Five contour levels are measured. Histograms show the relative occurrences of distance and angle values. The top histogram shows the occurrences of angles for chlorine, bromine, and iodine separately in bins of 10°, from 0°−180° relative to the sum of all data points. The right histogram shows the occurrences of distances for chlorine, bromine, and iodine separately in bins of 0.1 Å, from 2.0 Å - 4.0 Å relative to the sum of all data points.

Interaction angles >160° can be considered highly favorable. Further information and details of the individual contacts of the PDB scan can be found as an Excel Spreadsheet in the Supporting Information.

Hydrogen Bond Network Optimization and Calculation of Adduct Formation Energy. Based on restrictions of distance $d_{X\cdots O}$ and σ-hole angle $\alpha_{C-X\cdots O}$, we focused on 741 promising XB interactions in our subsequent analysis. The B-factor given for each atom in a PDB structure is a measure of uncertainty in the atom's position. B-factors of the interacting water molecule were compared to values of the interacting ligand and the surrounding protein binding site. If the value of the water molecule is more than 2.5 times higher than either the value of the ligand or atoms of the protein, the structure is inspected manually. Such high values indicate a high inaccuracy in the atom's position due to thermal motion. After manually inspecting the electron density, all structures were removed when clearly there was not enough electron density visible for either the water molecule or the ligand, thus, reducing the data set to 721 structures.

Furthermore, resolution of current X-ray crystallography does not enable the accurate detection of hydrogen atom positions. To address uncertainties and incorrectness of hydrogen positions, all present hydrogen atoms were removed. Extracted structures were protonated and hydrogen bond networks were initially optimized using Schrödinger's Protein Preparation Wizard, while all other atoms were kept frozen. The so retrieved structures were used as starting points for the subsequent QM optimization of the hydrogen bond network. Evaluation of the ligands in crystal structures of the PDB shows a large range of ligand sizes and chemical variety.[56,57] Often, aromatic or heteroaromatic scaffolds are decorated with halogen atoms. For simplicity and comparability, ligands were replaced with a simple model system of a corresponding halobenzene. Here, the halogen atom and the vector of the C−X bond of the rigid, optimized halobenzene are matched exactly onto the original ligand. The plane of the benzene ring is then optimally aligned onto the ligand's aromatic ring system. Since the halobenzene model system is much less tuned than typical ligands, we expect stronger interaction

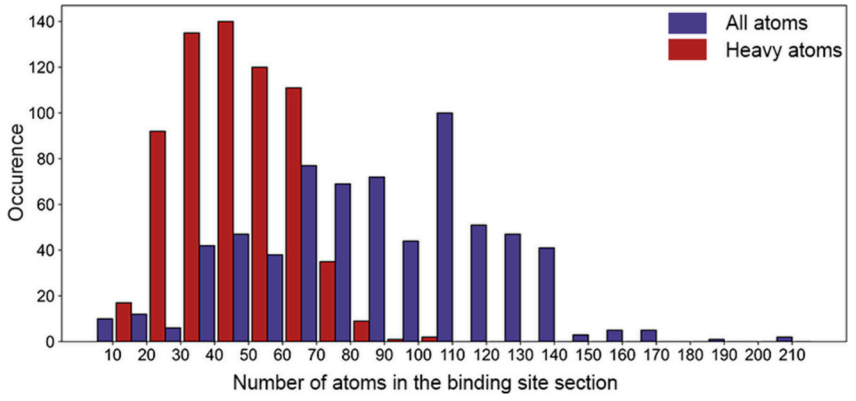


Figure 3. Number of atoms included in the binding site representations extracted from PDB structures for optimization of the hydrogen atom positions. Atom count is given in bins of 10. Occurrences are given for each bin. Blue bars show the total number of atoms including hydrogens in the binding site. Red bars show the number of heavy atoms.

https://doi.org/10.1021/acs.jctc.4c00834
J. Chem. Theory Comput. 2024, 20, 10716−10730

energies in regular ligand systems, where different scaffolds or electron withdrawing groups affect the strength of the σ-hole.[17] A total of 721 binding site representations were subjected to QM optimization of the hydrogen bond network. The geometry of a total of 652 structures successfully converged and was further analyzed. Figure 3 illustrates the distribution of binding site sizes in terms of heavy atom count and all atoms included in the binding site for hydrogen bond optimization. The size of the total binding site certainly has a crucial impact on the runtime and complexity of the optimization step.[58,59] Hence, the right balance between size, accuracy, calculation time and costs, the applied level of theory and the avoidance of convergence failures had to be chosen. Eventually, we optimized the hydrogen bond networks on a TPSS-D3/TZVP level of theory using TURBOMOLE.

To address any bias of overrepresentation in interaction geometries, only unique interactions within a PDB structure were considered. Therefore, interactions that occurred in multiple chains of the same PDB structure and turned out to be equal were neglected. Multiple interactions within the same chain addressing different water residues were kept. The final data set consists of 516 unique structures with 386 chlorine, 104 bromine, and 26 iodine interactions onto water.

Adduct Formation Energy Evaluation. We emulated the method applied for DYRK1a kinase (PDB: 8R8E) to the other PDB structures resulting from our selection process. To obtain the adduct formation energy and to assess the halogen bond strength, MP2/TZVPP single-point calculations of the individual interaction partners were conducted.

Table 3 shows the summary of the resulting mean energies and their standard deviations for each halogen interaction

Table 3. Mean Energies of the Halogen Interactions (in kJ/mol) Found in PDB Interaction Geometries

	All interactions			Only favorable		
	ΔE	N	SD	ΔE	N	SD
Chlorine	−1.41	386	5.31	−2.64	341	1.39
Bromine	−1.57	104	6.50	−3.24	88	1.59
Iodine	−4.22	26	5.58	−5.27	25	2.01

separately. Chlorine and bromine interactions show a mean energy of −1.41 kJ/mol, and −1.54 kJ/mol, respectively. Iodine interactions, however, show a mean energy of −4.22 kJ/mol for all interactions. It should be noted that the standard deviations are rather high and even some quite repulsive contacts are identified. Neglecting such positive energy values, and therefore unfavorable interactions, the mean energies are moderately improved to −2.64 kJ/mol, −3.24 kJ/mol, and −5.27 kJ/mol for chlorine, bromine, and iodine, respectively. The complete data set with all energies, distances, and angles can be found as an Excel Spreadsheet in the Supporting Information.The increasing interaction strength from chlorine to iodine correlates with the enhanced polarizability and atomic size of the halogens in terms of their van der Waals radii and, consequently, the larger σ-hole of the heavier halogens.[19]

To illustrate the diversity of binding geometries, we plotted the interaction distance $d_{X\cdots O}$ against the interaction angle $\alpha_{C-X\cdots O}$ in Figure 4. The best energy is highlighted as a diamond shaped data point. Each data point is colored by their respective adduct formation energy, according to the given color scale for the energy range from −10.0 to 0.0 kJ/mol.

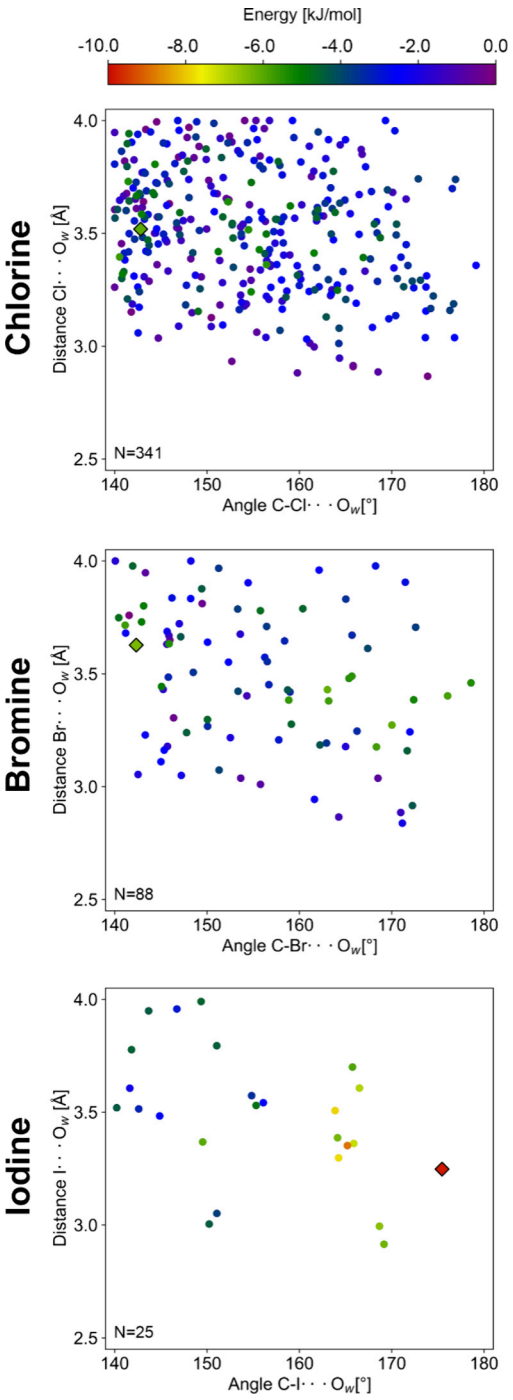


Figure 4. Distribution of favorable adduct formation energies for chlorine (N = 341), bromine (N = 88), and iodine interactions (N = 25). Geometries are characterized and differentiated by plotting the interaction angle $\alpha_{C-X\cdots O}$ against the interaction distance $d_{X\cdots O}$. Each data point is colored according to its adduct formation energy with respect to the given scale. The diamond shaped data point, featuring a black edging, emphasizes the best interaction geometry found.

Chlorine, with N = 341 interactions, shows a broad distribution of moderately negative energies. Bromine, consisting of N = 88 interactions, shows a similar spatial distribution and slightly more favorable interactions compared to chlorine. With only N = 25, iodine interactions are more scarce, but display the most favorable adduct formation energies of the three halogen elements. In stark contrast to

https://doi.org/10.1021/acs.jctc.4c00834
J. Chem. Theory Comput. 2024, 20, 10716−10730

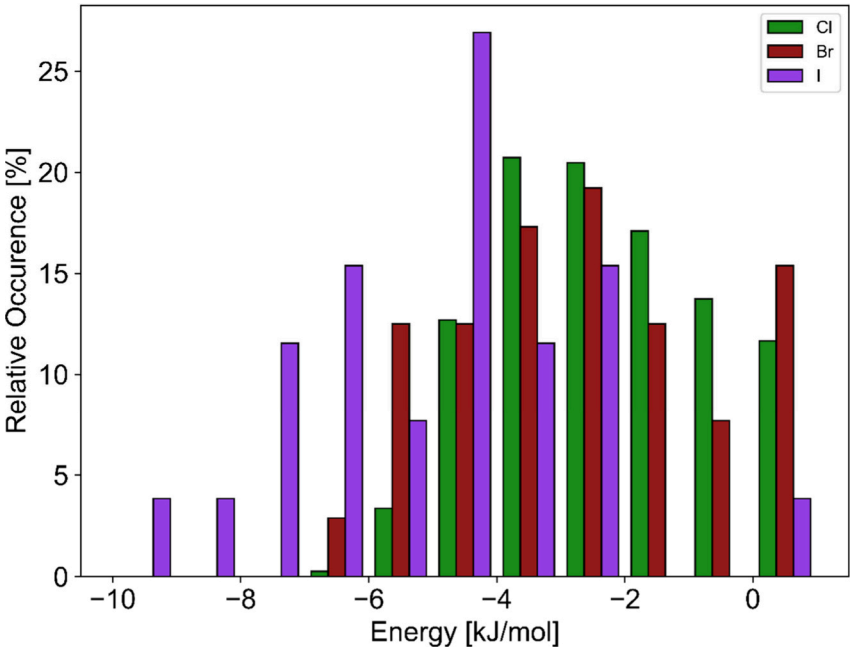


Figure 5. Histogram of relative occurrences of energy values for chlorine, bromine, and iodine interactions. Energy values are summed up from −10.0 to 0.0 in bins of −1.0 kJ/mol. An additional bin is added for values >0.0. Occurrences are reported relative to the sum of all data points of the respective halogen.

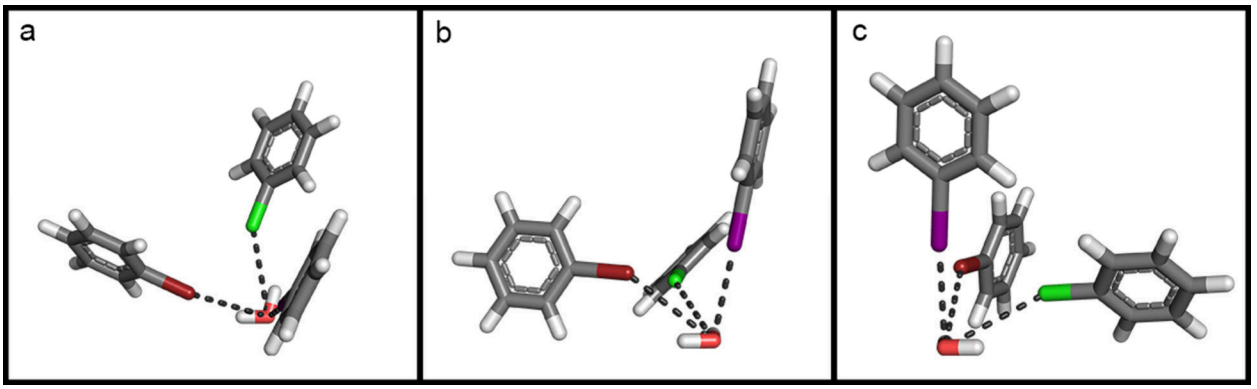


Figure 6. Spatial arrangement of best scoring interaction geometries for chlorobenzene, bromobenzene, and iodobenzene in complex with the addressed water molecule, extracted from PDB structures. The chlorine interaction (green) is found in 4UAL, the bromine interaction (dark red) in 5S41, and the iodine interaction (purple) in 5U84. The original ligands were replaced by the respective halobenzene as described in detail in the method section. (a−c) show different orientations of the same geometries. Dashed lines represent the X···O contact, allowing for a comparison with the expected directionality of a halogen bond. Figures were prepared with PyMOL.

the chlorine and bromine plots, in the case of iodine, a clear gradient of increasing adduct formation energies with higher σ-hole angles $\alpha_{C-X\cdots O}$ is easily evident.

Figure 5 shows the energy value distributions of adduct formation energies for chlorine, bromine, and iodine interactions, separately. Energy values are combined in bins of 1.0 kJ/mol from −10.0 to 0.0. An additional bin represents repulsive values >0.0 kJ/mol. The distribution reveals that iodine has a significant peak at around −6 kJ/mol, while also reaching most negative values of nearly −10 kJ/mol. Bromine and chlorine show a more uniform distribution across the energy range, with notable occurrences from −4 kJ/mol to −2 kJ/mol.

Iodine forms the strongest interaction with an energy of −9.83 kJ/mol, an interaction distance $d_{I\cdots O}$ = 3.25 Å, and an angle of $\alpha_{C-I\cdots O}$ = 175.5°. Weaker interaction energies are obtained for bromine and chlorine. The strongest interaction found for bromine is −6.34 kJ/mol with an interaction distance and angle of $d_{Br\cdots O}$ = 3.62 Å and $\alpha_{C-Br\cdots O}$ = 142.3°, respectively. For chlorine the most favorable energy is −6.02 kJ/mol with $d_{Cl\cdots O}$ = 3.52 Å and $\alpha_{C-Cl\cdots O}$ = 142.8°.

Figure 6 shows the respective geometry of each halogen interaction. Three different points of view are given to illustrate the spatial arrangement. Dashed lines illustrate the directionality of the formed halogen bond. The directivity of the σ-hole given by the C−I bond of the iodobenzene perpendicular to the plane formed by the water atoms clearly points to the lone pair of the oxygen atom. For the chlorine and bromine interaction the C−X vector shows an interesting pattern of a more "shifted, anti-parallel" direction in comparison to the O−H bond vector of the water molecule. This indicates that the halogen atom engages as a halogen bond donor with the oxygen atom, but can simultaneously act as a hydrogen bond acceptor. In this case, the hydrogen bond is formed to the

https://doi.org/10.1021/acs.jctc.4c00834
J. Chem. Theory Comput. 2024, 20, 10716−10730

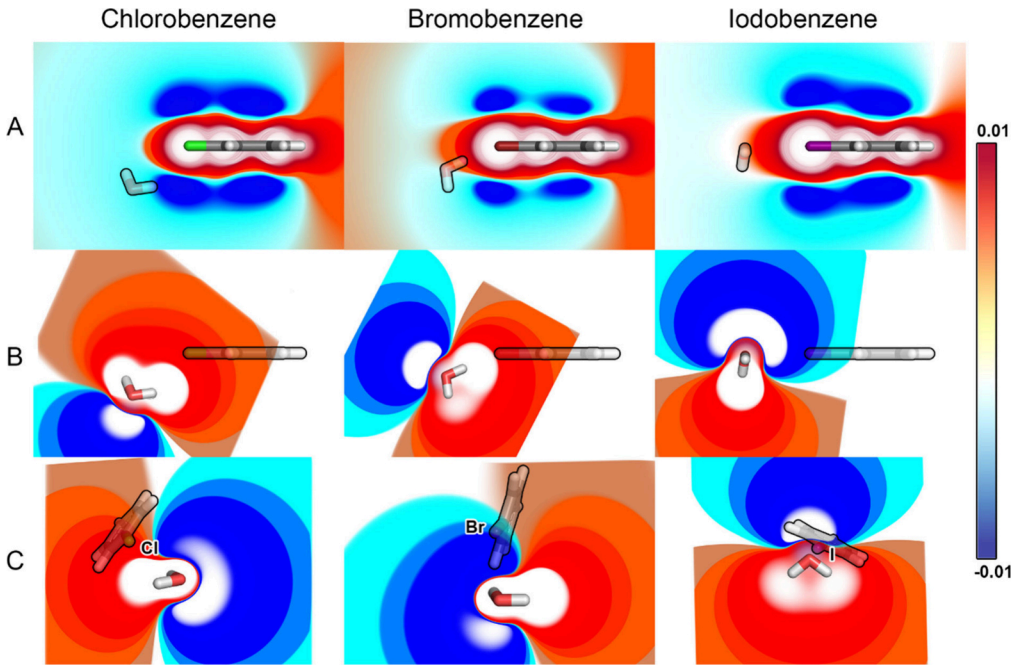


Figure 7. Depiction of different electrostatic potentials (ESP) of the top scoring geometries of chlorobenzene, bromobenzene, and iodobenzene in complex with the addressed water molecule. ESPs are calculated for each system separately, not for the complex. Positive ESPs at an energy of +0.01 au are colored in red, + 0.005 au in orange. Negative ESPs at an energy of −0.01 au are colored in blue, −0.005 au in cyan. Values at approximately 0.000 au, indicating the transition between positive and negative ESPs, are colored white. (A) ESPs shown for chlorobenzene, bromobenzene, and iodobenzene, illustrating the increasing σ-hole for: Cl < Br < I. A cutting plane through the C−X bond was used to show ESPs from the inside. For better visibility, the uncut halobenzene was rendered separately and superimposed later. In addition, a ghost molecule with a black outline highlights the position where the water molecule would be found in the interaction complex. (B) ESP of the water molecule in the same orientation as the respective ghost water molecule in (A). A cutting plane through the oxygen parallel to the point of view was used. The water molecule was rendered separately and superimposed later. A ghost molecule with a black outline illustrates the position of the corresponding halobenzene. (C) Selected orientations of the interaction geometries, chosen to visualize differences in the directivity of the C−X vector toward the water molecule. As in (B), a cutting plane through the oxygen atom and parallel to the point of view was used. The water molecule was rendered separately and superimposed later. A ghost molecule with a black outline illustrates the position of the corresponding halobenzene. Figures were prepared with PyMOL.

negatively charged belt in the equatorial position of the halogen atom (δ^- interactions). To clarify the shifted, antiparallel interaction profile, Figure S3 can be found in the Supporting Information showing each interaction separately in three different orientations. To illustrate the different interaction types based on matches and mismatches of the electrostatic potential (ESP) of each individual molecule, calculations on a regular grid around each system (chlorobenzene, bromobenzene, or iodobenzene) were conducted separately. Figure 7 A visualizes the ESPs of the three halobenzenes emphasizing the increased strength of the σ-hole for iodine over bromine, and chlorine. The water is depicted as a ghost molecule indicating its position and orienation in the respective top scoring geometry (Figure 6). Complementary to Figure 7 A, the ESP of the water molecule was calculated and visualized, while the halobenzenes of the respective interaction geometries are depicted as ghost molecules in the exact same orientation (Figure 7 B).

An additional orientation was selected to illustrate differences in the directivity of the C−X vector onto the water molecule (Figure 7 C). In the background, the halobenzene is shown again as ghost molecule. For better visibility, the halogen atom is labeled. From Figure 7 B and C, we identify clear differences between the three halogen atoms. While bromine and iodine clearly point toward the "lone pair" (blue region), chlorine points obviously past it. Iodine shows a clear overlap of the σ-hole with the negative potential caused by the "lone pair" of the water. Bromine still indicates a partial overlap of the σ-hole with the "lone pair", while also forming a "side-on" weak hydrogen bond.

Directionality, represented by the σ-hole angle $\alpha_{C-X\cdots O}$, is an important geometric feature for classifying a contact as a halogen bond, mediated through the σ-hole. In our PDB scan process, this parameter has been monitored and discussed as an important prerequisite for halogen bonding. However, as demonstrated in Figure 6 and Figure 7 for the best interactions identified for each element, the spatial arrangement of the halobenzene toward the "lone pairs" of the water molecule is similarly essential for an attractive σ-hole-based interaction. This can only be addressed after optimization of the proton positions of the water molecule in the context of the binding site. To investigate the energy-dependence of this spatial distribution of halogen atoms around the water molecule, we depict in Figure 8 the halogen atom positions with respect to a default orientation of the water molecule, differentiated by the type of halogen atom and the relative energy level of the interaction. All halogen atom positions are transformed into one hemisphere for better visibility. Energies were normalized by the best adduct formation energy of the respective halogen data set (geometries indicated by diamonds in Figure 4 and shown in Figure 6). The resulting range of normalized energies ranges between 0.0 for the start of attractive interactions and 1.0 for the most favorable interaction detected. Looking at the interaction geometries in the top range of 0.8 to 1.0, chlorine

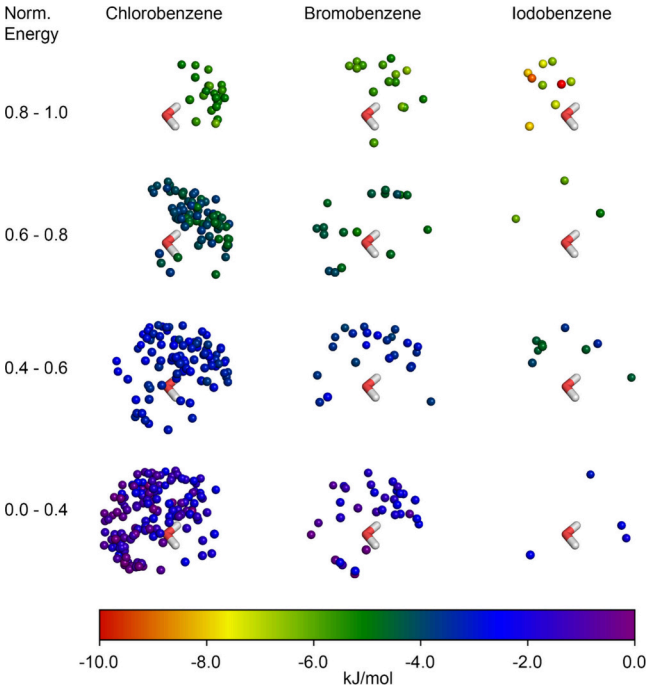


Figure 8. Halogen position overview of the initial halogen interaction geometry data set for chlorobenzene (N = 341), bromobenzene (N = 88), and iodobenzene (N = 25) in complex with water. Halogen positions are clustered in bins of normalized energies. Spheres are colored according to their respective energy bases on the given scale.

atom positions are always located in closer proximity to the hydrogen atoms than the oxygen atom. Bromine positions slightly shift to an area above the oxygen atom, whereas iodine positions clearly shift toward the oxygen atom. This pattern is significantly less pronounced for the weaker interactions (range 0.0−0.8). Since halogen atom positions alone do not give information about the directionality of the σ-hole, we also examined the spatial distribution of vectors of the C−X bond addressing the water molecule (Figure 9). Focusing on the best 20% of interaction geometries in terms of the adduct formation energy (normalized energies from 0.8 to 1.0), the figure shows the C−X vectors as sticks and the halogen atom positions as spheres. Each vector/sphere is colored with the corresponding adduct formation energy and exemplifies the spatial and directional preferences of the respective halogen interactions. Iodine vectors dominantly show a perpendicular σ-hole directionality toward the plane spanned by the water atoms, clearly addressing the lone pair of the oxygen atom. It has to be noted, that iodine also engages in such "shifted anti-parallel" interaction profiles. Considering the overall small amount of iodine interactions, four of the 25 geometries representing 16% of the interactions, display a spatial arrangement where hydrogen-bonding waters may still contribute to the interaction, albeit with weaker energies. While most bromine vectors still point toward the lone pairs, few interactions tend to shift into the previously mentioned "shifted, anti-parallel" orientation. In contrast, this orientation can be observed for most of the chlorine interactions. The vector distribution hints at a mixed interaction profile of both, halogen and hydrogen bonding. This finding mainly occurs among chlorine interactions and we suggest that it is due to the smaller size and extent of the σ-hole for chlorine, while the lateral negative electron density is more pronounced.

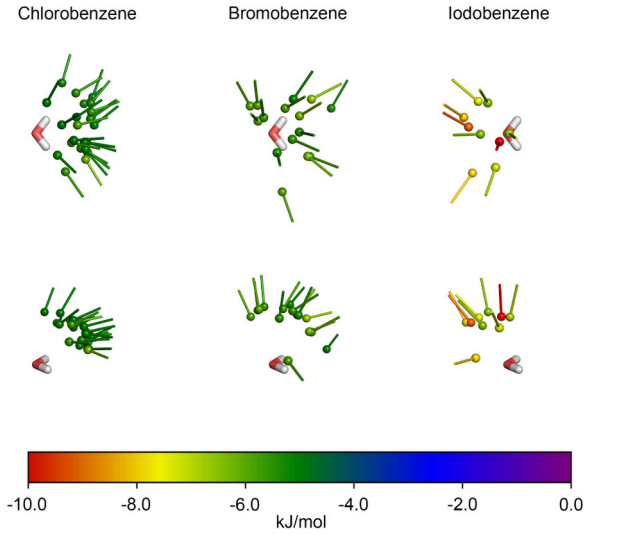


Figure 9. Depiction of the initial halogen interaction geometry data set for the best 20% of adduct formation energies for chloro-, bromo-, and iodobenzene in complex with water. Each sphere represents the position of the respective halogen, while vectors illustrate the C_{ar}−X bond. Spheres and vectors are colored according to the respective interaction energy based on the given scale.

Expansion to Virtual Data Sets through a Matched Molecular Pair Approach. While the analysis so far is based on experimental data, the size of the data sets is unbalanced due to a much higher preference for chlorinated ligands compared to brominated or iodinated ligands in the PDB. As a first step toward balancing this inequality, we have generated further geometries for virtual data set expansion from the original geometries by employing the complementary halobenzenes. Such a "gedankenexperiment", following the concept of virtual matched molecular pairs, can provide evidence, of whether simple halogen exchange can improve or diminish the interaction energy, while exactly conserving the binding mode. Based on different C−X bond lengths, there is, in principle, a halogen-centered and a scaffold-centered approach for this exchange. Either, in the halogen-centered approach, both halogen atoms are exactly matched onto each other, retaining interaction distance and angle, followed by exactly matching the C−X bond and the plane of the aromatic scaffold. Alternatively, in a scaffold centric approach, the aromatic scaffolds are matched onto each other, starting with the neighboring carbon atoms of the halogens, followed by aligning the C−X bond and the plane of the aromatic scaffold. While the latter method is useful from a molecular design perspective, it would lead to halogen···water contacts with not directly comparable geometries. Thus, the halogen-centered approach was employed herein. Based on the PDB results, there is a strong inequality of examples found for chlorine, bromine, and iodine. Through this data set expansion, we retrieve 516 examples for each halogen, strongly increasing the number of examples for iodine (+490), bromine (+412), and to a lesser extent for chlorine (+130).

Table 4 shows the mean energies for all interactions of chlorine (−1.4 kJ/mol), bromine (−1.58 kJ/mol), and iodine (−1.61 kJ/mol). Considering only favorable energies, values decrease to −2.6 kJ/mol, −3.34 kJ/mol, and −4.61 kJ/mol, respectively. The complete data set with all energies, distances, and angles can be found as an Excel Spreadsheet in the Supporting Information. Figure 10 shows the distribution of

Table 4. Mean Interaction Energies of the Virtual SAR Study for All Three Halobenzenes (in kJ/mol)

	All interactions (N = 516)		Only favorable	
	ΔE	SD	ΔE	SD
Chlorine	−1.40	5.23	−2.60 (N = 456)	1.36
Bromine	−1.58	6.99	−3.34 (N = 446)	1.6
Iodine	−1.61	10.21	−4.61 (N = 422)	2.32

interaction angles $\alpha_{C-X\cdots O}$ and interaction distance $d_{X\cdots O}$ of the virtually complemented data sets. Triangular shaped data points derive from the original data set, while circular data points represent the virtually exchanged, and therefore, newly added data points. For chlorine and iodine, the best interaction energies remain the same as before, while for bromine, a newly top-scoring interaction energy was identified. Since iodine interactions in general are sparsely represented in crystal structures of the PDB, it is not surprising that a huge number of virtual geometries are added through our halogen exchange. Still, it is noteworthy that a quite significant number of stronger interactions are found within a distance range $d_{I\cdots O}$ of 3.0 Å to 3.5 Å and an angle range $\alpha_{C-I\cdots O}$ of 160° to 180°.

To showcase the benefit that halogen exchange can have on the adduct formation energy of an interaction, we focus here on a few examples. Figure 11 shows pairwise halogen exchanges and the corresponding increase in energy. Exchanging iodine to bromine in the interaction geometry found in 5S40, yields an energy increase of nearly 60% from −2.09 kJ/mol to −3.57 kJ/mol, while exchanging to chlorine, the energy is even doubled to −4.22 kJ/mol. In 5S40, the iodine's σ-hole does not address the lone pair of the water oxygen atom, but shows a shifted antiparallel interaction pattern. Hence, in this binding situation, replacing iodine by chlorine and bromine seems intuitive, because these halogens will benefit more strongly from the halogen and hydrogen interactions of such a shifted antiparallel binding mode.

For chlorine, the exchange to one of the heavier halogens, can be seen as a natural halogen bond tuning effect, due to the increase in σ-hole size and strength. Given the interaction geometry found in 6DUV, the chlorine's σ-hole clearly points to the negatively charged lone pair. As a result, the exchange to bromine increases the interaction energy from −2.75 kJ/mol to −5.14 kJ/mol, while an exchange to iodine will further strengthen the interaction energy to −8.85 kJ/mol, ranking among the most favorable interactions observed.

Based on its intermediate position between chlorine and iodine, when replacing bromine, two distinct cases need to be differentiated: Exchange to chlorine in the binding mode of 2QLQ (no halogen bond) increases the energy from −2.42 kJ/mol to −3.91 kJ/mol. Replacing bromine with iodine in the geometry of 5HEX (halogen bond) increases the interaction energy from −5.43 kJ/mol to −8.67 kJ/mol, providing again an example of one of the highest-scoring geometries.

As previously examined for the initial halogen interaction data sets in the PDB (Figure 8 and Figure 9), spatial distribution and directionality of the C−X bond vectors for these enhanced data sets are depicted in Figure 12 and Figure 13, respectively. While Figure 13 focuses only on the 20% best interaction geometries, Figure S1 in the Supporting Information (complemented by an individual Figure_S1_Cl.pse, Figure_S1_Br.pse, and Figure_S1_I.pse file) depicts the entire data set as shown in Figure 12.

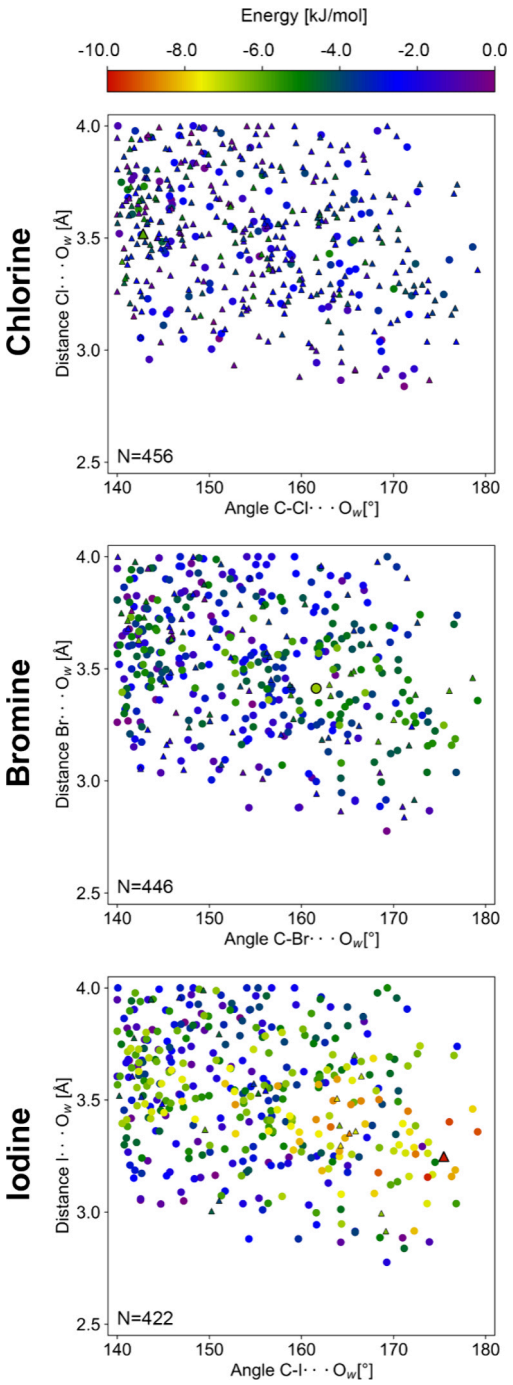


Figure 10. Distribution of adduct formation energies for chlorobenzene, bromobenzene, and iodobenzene molecules with respect to interaction distance $d_{X\cdots O}$ and interaction angle $\alpha_{C-X\cdots O}$. Each data point is colored according to the respective interaction energy based on the given scale. Triangular shaped data points with black stroke are derived from the original PDB data set, while circled data points are derived as the complementary data set through halogen exchange. The best interaction energy is indicated by a larger data point with a black stroke.

What has been observed anecdotically in Figure 11, appears as a more obvious trend in Figure 12 (compared to Figure 8) and Figure 13 (compared to Figure 9) based on the enhanced number of representative structures. Looking at the best 20% of the interaction geometries (normalized energy: 0.8−1.0), preferred chlorine atom positions were observed consistently

https://doi.org/10.1021/acs.jctc.4c00834
J. Chem. Theory Comput. 2024, 20, 10716−10730

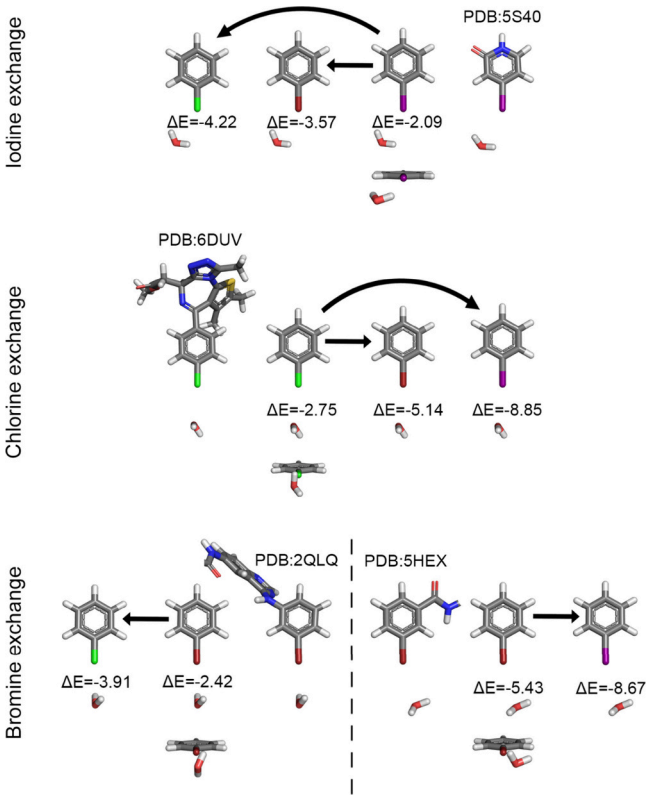


Figure 11. Examples of pairwise halobenzene exchanges yielding an energy gain. The halogen position remains the same throughout the exchange.

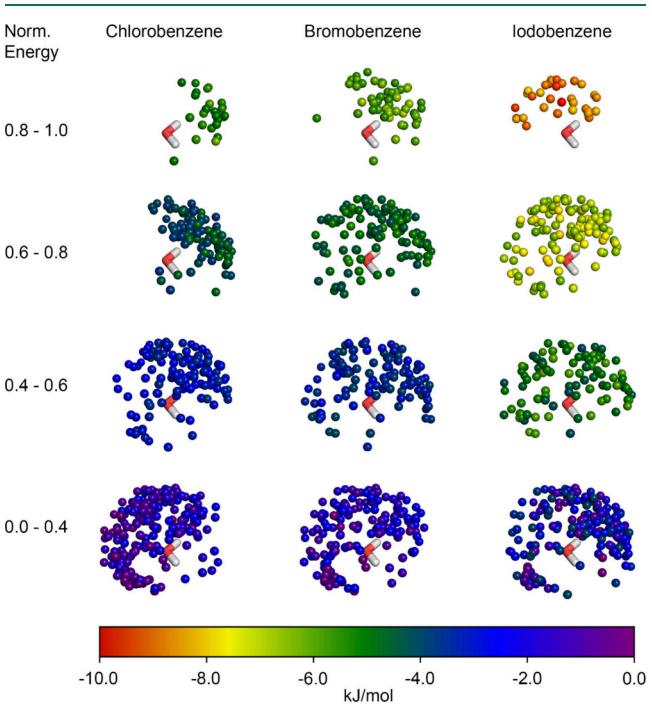


Figure 12. Halogen position overview of the expanded data set of interaction geometries for chlorobenzene, bromobenzene, and iodobenzene in complex with water. Halogen positions are clustered in bins of normalized energies. Spheres are colored according to their respective energy bases on the given scale.

in a region above the hydrogen atoms. In combination with their C−X bond vectors (Figure 13), this emphasizes that

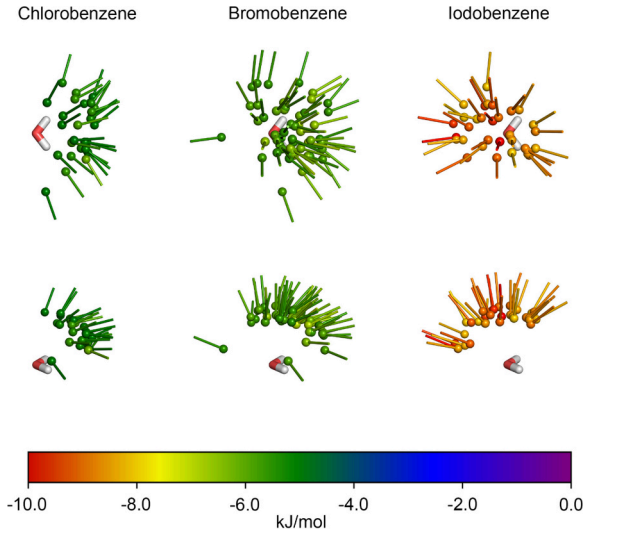


Figure 13. Depiction of the virtually expanded halogen interaction geometry data set for the best 20% of adduct formation energies for chlorobenzene, bromobenzene, and iodobenzene in complex with water. Each sphere represents the position of the respective halogen, while vectors illustrate the C_{ar}−X bond. Spheres and vectors are colored according to the respective interaction energy based on the given scale.

chlorine tends to adopt shifted antiparallel geometries, engaging partially in halogen and hydrogen bonding. Bromine atom positions slightly shift to an area above the oxygen atom, whereas iodine atoms are located clearly above the oxygen atom, approaching its lone pairs. Vectors of the C−Br bond are distributed more evenly around the full hemisphere (Figure 13). A larger fraction compared to chlorine shows the typical σ-hole directionality. Eventually, most C−I bond vectors and iodine positions feature a preferential σ-hole directionality, indicating halogen bonding with the water oxygen atom.

CONCLUSION AND OUTLOOK

The herein presented evaluation of halogen−water interactions in the binding site of proteins illustrated that addressing a water molecule through halogen bonding can, in principle, yield a beneficial contribution to protein−ligand binding. We have started to elucidate, whether it is just a "drop in the ocean" or can and should be harnessed in molecular design.

We identified an interesting halogen−water interaction in our crystal structure of 2-cyclopentyl-7-iodo-1*H*-indole-3-carbonitrile binding to DYRK1a kinase (PDB: 8R8E) and carefully evaluated the binding situation. Subsequently, a PDB scan, followed by QM calculations of interaction geometries was conducted. Our results show that the interaction with water molecules can be highly versatile depending on the halogen, and of course, the binding situation. The strongest interaction energy of −9.83 kJ/mol is based on a good XB geometry of an iodinated ligand. In previous studies targeting the backbone carbonyl,[19] the sulfur of methionine,[21] and the carboxamide group of asparagine or glutamine by halogen bonds,[23] optimal interactions of close to −20 kJ/mol were found.[45] For XB on the carboxylate group of aspartate or glutamate,[23] interaction energies even in the range of −56 kJ/mol were observed, due to the impact of the negative charge of the anionic carboxylate function. In comparison to such contributions to molecular recognition, the resulting interaction energies with water appear underwhelming at first sight.

https://doi.org/10.1021/acs.jctc.4c00834
J. Chem. Theory Comput. 2024, 20, 10716−10730

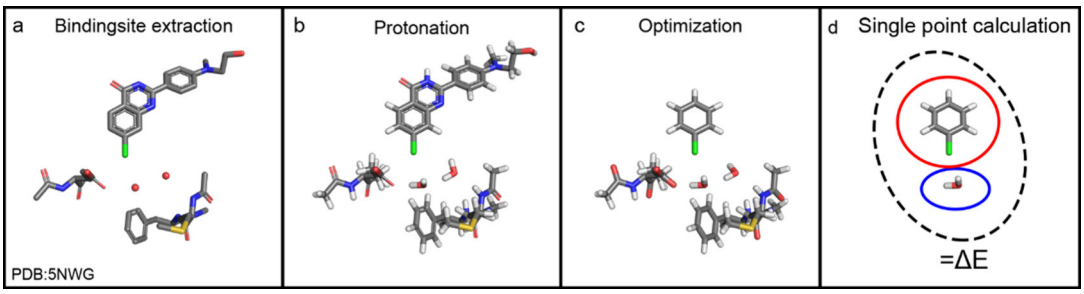


Figure 14. Overview of the workflow to extract interaction geometries from binding sites in PDB crystal structures. (a) Example binding site selection (PDB: 5NWG) containing the ligand, the addressed water molecule, and further surrounding amino acids and water molecules within 4 Å of the ligand and water of interest. Loose ends of amino acids are extended by two bonds to avoid charges at the backbone. (b) Protonation of the extracted binding site section using Schrödinger's Protein Preparation Wizard module. (c) The original ligand is replaced by the corresponding halobenzene. Optimization of the hydrogen network on a TPSS(D3)/TZVP level of theory using TURBOMOLE. (d) Adduct formation energies are calculated on a MP2/TZVPP level of theory for extracted interaction geometries of halobenzene (red), water molecule (blue), and of both in complex (dashed black).

Still, stronger interactions would consequently increase the costs of desolvation accordingly. Thus, replacing water molecules from the unbound state of the ligand with a "more favorable" interaction partner in the binding site, would become significantly more difficult to achieve, if not impossible.

Obviously, the other elemental contribution to the free energy of binding is entropy. As often evidenced in drug discovery, a certain enthalpy/entropy compensation can occur. While our focus in this study is not an assessment of entropy, it can still be useful to contemplate our findings with respect to entropic effects. From Figure 12, we deduce that point clouds representing a diverse set of geometries, but showing similar color (i.e., energy), indicate more mobility of the ligand and, thus, more degrees of freedom for the ligand−water complex. Such distributions are more clearly found in the case of chlorine, than bromine or iodine. We already discussed that chlorine−water interactions are often preferring a shifted antiparallel binding mode, showing intermediate states between clear halogen or clear hydrogen bonding. In contrast to the high directionality of halogen bonds, such intermediate interactions are significantly more flexible. Both aspects suggest, that for chlorine−water contacts, a lesser enthalpic interaction strength could be (partially) compensated by a stronger entropic contribution to the free energy of binding. The opposite situation is found for iodine, where a broader range of interaction energies occur, and the best interaction geometries are restricted to an ideal placement, yielding a quite directional halogen bond interaction through the σ-hole. Hence, a stronger enthalpic contribution to the free energy of binding could be (partially) compensated by a weaker entropic contribution.

The situation may change significantly, when the water molecule is not highly flexible, as in "bulk water", but is engaging in up to three hydrogen bonds as an "interstitial" water toward the binding site. In case that the water molecule is oriented to facilitate accepting an additional halogen bond through one of its lone pairs, while not having any further restraining effect on the water molecule, the enthalpic gain of the formed halogen bond could improve the affinity of the ligand compared to a more flexible water molecule. Thus, "interstitial" water molecules, identified as stable parts of the binding site by having quite similar B-factors as compared to the protein environment, might be prime targets for halogen bonding. This effect could be likewise relevant for inducing selectivity, because an assembly of identical amino acids in two quite homologous binding sites, could still differ enough in its spatial arrangement to bind an "interstitial" water molecule more tightly or loosely.

Based on these considerations, we suggest that it will largely depend on where the "drop" is located in the "ocean" to determine, whether it is worth targeting this water molecule. Hence, in the future, more systematic studies will be required to assist this initial analysis. Likewise, the combination of crystal structures of higher resolution, considerate crystallographers and detailed analysis of the binding modes of waters using quantum chemistry could provide further valuable examples of water molecules that make a difference in drug discovery.

MATERIALS AND METHODS

DYRK1a Crystallization. Crystallization was performed using the sitting drop vapor diffusion method at 4 °C as previously described.[60] The protein containing the His-Tag was concentrated to 12 mg/mL and AMP-PCP was added to a final concentration of 1−2 mM. The reservoir solution consisted of 100 mM TRIS, 100 mM Li_2SO_4 and 34% (v/v) PEG300 at pH = 8.5. Protein (4 μL) was mixed with the solution (2 μL) and crystals grew within a week. Streak seeding was performed to enhance crystal growth and quality. 2-Cyclopentyl-7-iodo-1*H*-indole-3-carbonitrile (Compound YOX) was added to a fresh drop of reservoir solution containing 5% DMSO to a final concentration of 1 mM and crystals were soaked for about 24 h.

Data Collection and Refinement. Diffraction data was collected at the SLS beamline X06DA. Data reduction was performed using XDS.[61] Initial phases were obtained using PHASER[62] included in the CCP4 suite[63] and using the searchmodel 2VX3.[64] Structure and phase improvement was performed in multiple cycles of manual building in Coot[65] and refinement using PHENIX.[66] Restraints for 2-cyclopentyl-7-iodo-1*H*-indole-3-carbonitrile were generated using AceDRG[67] as part of the CCP4 suite.

Analysis of Protein−Ligand Complexes in the Protein DataBank (PDB). A PDB scan was conducted using a custom Python/PyMOL script. Alternative conformations, metal ions and hydrogen atoms were removed from the structure. PDB structures containing halogenated ligands (chlorine, bromine, iodine) with the halogen atom connected to an aromatic ring system and a ligand size of six or more heavy atoms were

https://doi.org/10.1021/acs.jctc.4c00834
J. Chem. Theory Comput. 2024, 20, 10716−10730

considered for further analysis. Water residues within 4 Å of the halogen atom are retained. Structures were selected only if water and halogen atom show occupancies of 1.0. The B-factor of the relevant water molecules were compared to the average B-factors of the ligand and the surrounding protein residues.

Binding Site Selection and Hydrogen Bond Network Optimization. Interaction geometries between halogenated ligands and water molecules were extracted from the PDB structures. Additionally, all surrounding amino acid residues and other water molecules within 4 Å of the water molecule of interest and the halogen atom were included into the binding site section. Loose ends of amino acid residues were extended by two bonds including the backbone carbonyl and C_α of the following amino acid to avoid potential charges at the protein backbone (Figure 14a). Extracted binding site sections were protonated using the Protein Preparation Wizard module of Schrödinger suite version 2021−1 (Figure 14b). Additional parameters were used to fix heavy atoms and minimize sampled hydrogen atoms using the default force field.

We replaced each ligand with a halobenzene (chloro-, bromo-, iodobenzene) by matching the halogen atoms, the C−X bond vector and the planes of the (hetero)aromatic systems. Separate geometry optimization of the halobenzenes was done at the MP2-level of theory using TURBOMOLE 7.4.1 with a triple-ζ basis set (def2-TZVPP). Calculations were done in combination with the resolution of identity (RI) technique and the frozen core approximation. Frozen core orbitals were defined using default settings, where orbitals with energies below −3.0 au are considered core orbitals. SCF convergence criterion was increased to 10^{-8} hartree. Relativistic effects for iodine were considered by an effective core potential (ECP).[68−75] Hydrogen bond networks were optimized at the TPSS(D3) level of theory with a triple-ζ basis set (def2-TZVP) using TURBOMOLE 7.4.1 (Figure 14c). Calculations were done in combination with the resolution of identity (RI) technique. Heavy atoms were kept frozen. The TPSS functional was augmented with an empirical dispersion correction as proposed by Grimme,[51] which is indicated by adding "(D3)" to the name. The SCF convergence criterion was increased to 10^{-8} hartree.

Calculation of Adduct Formation Energies and Halogen Exchange. After the QM optimization of hydrogen atom positions, the halobenzene and the addressed water molecule were extracted from its binding site section (Figure 14d). Single point calculations of the halobenzene, the water molecule and of the complex of both were carried out at the MP2 level of theory with a triple-ζ basis set (def2-TZVPP). Furthermore, in each structure the halobenzene was exchanged with both other halobenzenes (e.g., chlorobenzene into bromobenzene or iodobenzene), keeping the position of the halogen, the C−X bond vector and the plane of the benzene ring identical. This systematic replacement was performed in this manner to allow for a direct comparison between the different halogen atoms for exactly the same interaction geometry. Adduct formation energies were calculated as

$$\Delta E = E_{complex} - (E_{ligand} + E_{water}) \quad (1)$$

and reported as kJ/mol.

■ ASSOCIATED CONTENT

Data Availability Statement

PyMOL is an open-source software maintained and distributed by Schrödinger. There is an open-source version of PyMOL available at: https://github.com/schrodinger/pymol-open-source. Python is an open-source programming language available and downloadable from https://www.python.org/. All crystal structure data and material used in this study is available free of charge at https://www.rcsb.org/. Detailed results of the PDB scan and of the QM-based evaluation of halogen−water interaction energies derived from the PDB scan results are provided in spreadsheet format (.xlsx) as Supporting Information. TURBOMOLE is a purchasable software maintained and distributed by the *TURBOMOLE GmbH*. Demo versions are available at https://www.turbomole.org/. The licensed software was provided to us by the bwHPC Cluster JUSTUS2. PDB-Code: 8R8E, Chemical ID in PDB structure: YOX, Residue ID: 501. Authors will release the atomic coordinates and experimental data upon article publication. DYRK1a protein structure in complex with 2-cyclopentyl-7-iodo-1*H*-indole-3-carbonitrile was deposited in the PDB under the accession code 8R8E. This material is available free of charge via the Internet at https://www.rcsb.org/.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jctc.4c00834.

Additional details, figures, and information on data collection and refinement of the crystal structure, of the PDB scan process, the energy evaluation, and interaction geometries (PDF)

PDB scan results (XLSX)

Halogen−water interaction energies (XLSX)

PyMOL sessions of the individual halogen position data sets, additional data and details to Figures 12 and 13 (ZIP)

■ AUTHOR INFORMATION

Corresponding Authors

Markus O. Zimmermann − *Laboratory for Molecular Design & Pharmaceutical Biophysics, Institute of Pharmaceutical Sciences, Department of Pharmacy and Biochemistry and Interfaculty Institute for Biomedical Informatics (IBMI), Eberhard Karls Universität Tübingen, 72076 Tübingen, Germany;* orcid.org/0000-0001-6115-8248; Email: m.zimmermann@uni-tuebingen.de

Frank M. Boeckler − *Laboratory for Molecular Design & Pharmaceutical Biophysics, Institute of Pharmaceutical Sciences, Department of Pharmacy and Biochemistry and Interfaculty Institute for Biomedical Informatics (IBMI), Eberhard Karls Universität Tübingen, 72076 Tübingen, Germany;* orcid.org/0000-0001-8738-6716; Email: frank.boeckler@uni-tuebingen.de

Authors

Marc U. Engelhardt − *Laboratory for Molecular Design & Pharmaceutical Biophysics, Institute of Pharmaceutical Sciences, Department of Pharmacy and Biochemistry, Eberhard Karls Universität Tübingen, 72076 Tübingen, Germany;* orcid.org/0009-0007-9152-8538

Marcel Dammann − *Laboratory for Molecular Design & Pharmaceutical Biophysics, Institute of Pharmaceutical*

Sciences, Department of Pharmacy and Biochemistry, Eberhard Karls Universität Tübingen, 72076 Tübingen, Germany; Present Address: M.D., Catalent Clinical Supply Services, 73614 Schorndorf, Germany

Jason Stahlecker − *Laboratory for Molecular Design & Pharmaceutical Biophysics, Institute of Pharmaceutical Sciences, Department of Pharmacy and Biochemistry, Eberhard Karls Universität Tübingen, 72076 Tübingen, Germany;* orcid.org/0000-0002-0044-4037

Antti Poso − *School of Pharmacy, Faculty of Health Sciences, University of Eastern Finland, 70211 Kuopio, Finland; Institute of Pharmaceutical Sciences, Pharmaceutical/Medicinal Chemistry and Tübingen Center for Academic Drug Discovery & Development (TüCAD2), Eberhard Karls Universität Tübingen, 72076 Tübingen, Germany;* orcid.org/0000-0003-4196-4204

Thales Kronenberger − *School of Pharmacy, Faculty of Health Sciences, University of Eastern Finland, 70211 Kuopio, Finland; Institute of Pharmaceutical Sciences, Pharmaceutical/Medicinal Chemistry and Tübingen Center for Academic Drug Discovery & Development (TüCAD2), Eberhard Karls Universität Tübingen, 72076 Tübingen, Germany; Excellence Cluster "Controlling Microbes to Fight Infections" (CMFI), 72076 Tübingen, Germany; Interfaculty Institute of Microbiology and Infection Medicine (IMIT), University of Tübingen, 72076 Tübingen, Germany; Partner-site Tübingen, German Center for Infection Research (DZIF), 72076 Tübingen, Germany*

Conrad Kunick − *Institute for Medicinal and Pharmaceutical Chemistry, Technische Universität Braunschweig, 38106 Braunschweig, Germany*

Thilo Stehle − *Interfaculty Institute of Biochemistry, Eberhard Karls Universität Tübingen, 72076 Tübingen, Germany*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jctc.4c00834

Author Contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript. F.M.B. and M.O.Z. envisioned the research. M.D. and J.S. prepared the protein, conducted protein crystallization experiments, and performed data reduction and structure refinement. C.K. reviewed the manuscript and provided 2-cyclopentyl-7-iodo-1*H*-indole-3-carbonitrile. M.U.E. performed all QM calculations, wrote applications for PDB analysis, conducted PDB analysis and prepared the corresponding visualizations. M.O.Z. contributed to developing the computational strategy. T.K. and A.P. helped refine the concept and provided comments on the manuscript. M.U.E. prepared the original draft. M.U.E., M.O.Z., and F.M.B. reviewed, edited, and finalized the manuscript.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors acknowledge support from the state of Baden−Württemberg through bwHPC and the German Research Foundation (DFG) through Grant No. INST 40/575-1 FUGG (JUSTUS 2 cluster). In addition, support is acknowledged from the High Performance and Cloud Computing Group at the Zentrum für Datenverarbeitung of the University of Tübingen and the German Research Foundation (DFG) through Grant No. INST 37/935-1 FUGG (BinAC cluster). T.K. is funded by the fortune initiative and from TüCAD2 and CMFI. TüCAD2 and CMFI are funded by the Federal Ministry of Education and Research (BMBF) and the Baden-Württemberg Ministry of Science as part of the Excellence Strategy of the German Federal and State Governments. The authors wish to acknowledge CSC − IT Center for Science, Finland, for the very generous computational resources.

■ ABBREVIATIONS

XB, halogen bond; X, halogen; QM, quantum mechanical; TPSS, Tao, Perdew, Staroverov, and Scuseria exchange functional; D3, dispersion correction type 3; TZVP, valence triple-ζ polarization; MP2, Møller−Plesset perturbation method order 2; TZVPP, valence triple-ζ with two sets of polarization functions; PDB, Protein Data Bank; ESP, electrostatic surface potential.

■ REFERENCES

(1) Bissantz, C.; Kuhn, B.; Stahl, M. A Medicinal Chemist's Guide to Molecular Interactions. *J. Med. Chem.* **2010**, *53* (14), 5061−5084.

(2) Politzer, P.; Murray, J. S.; Clark, T. Halogen bonding and other σ-hole interactions: a perspective. *Phys. Chem. Chem. Phys.* **2013**, *15* (27), 11178−11189.

(3) Andrea, R. V.; P, S. H. The Role of Halogen Bonding in Inhibitor Recognition and Binding by Protein Kinases. *Curr. Top. Med. Chem.* **2007**, *7* (14), 1336−1348.

(4) Auffinger, P.; Hays, F. A.; Westhof, E.; Ho, P. S. Halogen bonds in biological molecules. *Proc. Natl. Acad. Sci. U. S. A.* **2004**, *101* (48), 16789−16794.

(5) Cavallo, G.; Metrangolo, P.; Milani, R.; Pilati, T.; Priimagi, A.; Resnati, G.; Terraneo, G. The Halogen Bond. *Chem. Rev.* **2016**, *116* (4), 2478−2601.

(6) Erdélyi, M. Halogen bonding in solution. *Chem. Soc. Rev.* **2012**, *41* (9), 3547−3557.

(7) Shinada, N. K.; de Brevern, A. G.; Schmidtke, P. Halogens in Protein−Ligand Binding Mechanism: A Structural Perspective. *J. Med. Chem.* **2019**, *62* (21), 9341−9356.

(8) Hernandes, Z. M.; Cavalcanti, T. S. M.; Moreira, M. D. R.; de Azevedo Junior, F. W.; Leite, L. A. C. Halogen Atoms in the Modern Medicinal Chemistry: Hints for the Drug Design. *Curr. Drug Targets* **2010**, *11* (3), 303−314.

(9) Rowe, R. K.; Ho, P. S. Relationships between hydrogen bonds and halogen bonds in biological systems. *Acta Crystallogr. Sect. B: Struct. Sci.* **2017**, *73* (2), 255−264.

(10) Voth, A. R.; Khuu, P.; Oishi, K.; Ho, P. S. Halogen bonds as orthogonal molecular interactions to hydrogen bonds. *Nat. Chem.* **2009**, *1* (1), 74−79.

(11) Xu, Z.; Yang, Z.; Liu, Y.; Lu, Y.; Chen, K.; Zhu, W. Halogen Bond: Its Role beyond Drug−Target Binding Affinity for Drug Discovery and Development. *J. Chem. Inf. Model.* **2014**, *54* (1), 69−78.

(12) Zhu, Z.; Xu, Z.; Zhu, W. Interaction Nature and Computational Methods for Halogen Bonding: A Perspective. *J. Chem. Inf. Model.* **2020**, *60* (6), 2683−2696.

(13) Matter, H.; Nazaré, M.; Güssregen, S.; Will, D. W.; Schreuder, H.; Bauer, A.; Urmann, M.; Ritter, K.; Wagner, M.; Wehner, V. Evidence for C-Cl/C-Br···π Interactions as an Important Contribution to Protein−Ligand Binding Affinity. *Angew. Chem., Int. Ed.* **2009**, *48* (16), 2911−2916.

(14) Clark, T.; Hennemann, M.; Murray, J. S.; Politzer, P. Halogen bonding: the σ-hole. *J. Mol. Model.* **2007**, *13* (2), 291−296.

(15) Desiraju, G. R.; Ho, P. S.; Kloo, L.; Legon, A. C.; Marquardt, R.; Metrangolo, P.; Politzer, P.; Resnati, G.; Rissanen, K. Definition of the halogen bond (IUPAC Recommendations 2013). *Pure Appl. Chem.* **2013**, *85* (8), 1711−1713.

(16) Sedlak, R.; Kolář, M. H.; Hobza, P. Polar Flattening and the Strength of Halogen Bonding. *J. Chem. Theory Comput.* **2015**, *11* (10), 4727−4732.
(17) Lange, A.; Heidrich, J.; Zimmermann, M. O.; Exner, T. E.; Boeckler, F. M. Scaffold Effects on Halogen Bonding Strength. *J. Chem. Inf. Model.* **2019**, *59* (2), 885−894.
(18) Sakai, T.; Torii, H. Substituent Effect and Its Halogen-Atom Dependence of Halogen Bonding Viewed through Electron Density Changes. *Chem. Asian J.* **2023**, *18* (3), No. e202201196.
(19) Wilcken, R.; Zimmermann, M. O.; Lange, A.; Zahn, S.; Boeckler, F. M. Using halogen bonds to address the protein backbone: a systematic evaluation. *J. Comput. Aided Mol. Des.* **2012**, *26* (8), 935−945.
(20) Zimmermann, M. O.; Boeckler, F. M. Targeting the protein backbone with aryl halides: systematic comparison of halogen bonding and $\pi\cdots\pi$ interactions using N-methylacetamide. *MedChemComm* **2016**, *7* (3), 500−505.
(21) Wilcken, R.; Zimmermann, M. O.; Lange, A.; Zahn, S.; Kirchner, B.; Boeckler, F. M. Addressing Methionine in Molecular Design through Directed Sulfur−Halogen Bonds. *J. Chem. Theory Comput.* **2011**, *7* (7), 2307−2315.
(22) Lange, A.; Zimmermann, M. O.; Wilcken, R.; Zahn, S.; Boeckler, F. M. Targeting Histidine Side Chains in Molecular Design through Nitrogen−Halogen Bonds. *J. Chem. Inf. Model.* **2013**, *53* (12), 3178−3189.
(23) Zimmermann, M. O.; Lange, A.; Zahn, S.; Exner, T. E.; Boeckler, F. M. Using Surface Scans for the Evaluation of Halogen Bonds toward the Side Chains of Aspartate, Asparagine, Glutamate, and Glutamine. *J. Chem. Inf. Model.* **2016**, *56* (7), 1373−1383.
(24) Scholfield, M. R.; Zanden, C. M. V.; Carter, M.; Ho, P. S. Halogen bonding (X-bonding): A biological perspective. *Protein Sci.* **2013**, *22* (2), 139−152.
(25) Shah, M. B.; Liu, J.; Zhang, Q.; Stout, C. D.; Halpert, J. R. Halogen−π Interactions in the Cytochrome P450 Active Site: Structural Insights into Human CYP2B6 Substrate Selectivity. *ACS Chem. Biol.* **2017**, *12* (5), 1204−1210.
(26) Zhou, P.; Lv, J.; Zou, J.; Tian, F.; Shang, Z. Halogen−water−hydrogen bridges in biomolecules. *J. Struct. Biol.* **2010**, *169* (2), 172−182.
(27) Barillari, C.; Taylor, J.; Viner, R.; Essex, J. W. Classification of Water Molecules in Protein Binding Sites. *J. Am. Chem. Soc.* **2007**, *129* (9), 2577−2587.
(28) Geschwindner, S.; Ulander, J. The current impact of water thermodynamics for small-molecule drug discovery. *Expert Opin. Drug Dis.* **2019**, *14* (12), 1221−1225.
(29) Ladbury, J. E. Just add water! The effect of water on the specificity of protein-ligand binding sites and its potential application to drug design. *Chem. Biol.* **1996**, *3* (12), 973−980.
(30) Poornima, C. S.; Dean, P. M. Hydration in drug design. 1. Multiple hydrogen-bonding features of water molecules in mediating protein-ligand interactions. *J. Comput. Aided Mol. Des.* **1995**, *9* (6), 500−512.
(31) Poornima, C. S.; Dean, P. M. Hydration in drug design. 2. Influence of local site surface shape on water binding. *J. Comput. Aided Mol. Des.* **1995**, *9* (6), 513−520.
(32) Rudling, A.; Orro, A.; Carlsson, J. Prediction of Ordered Water Molecules in Protein Binding Sites from Molecular Dynamics Simulations: The Impact of Ligand Binding on Hydration Networks. *J. Chem. Inf. Model.* **2018**, *58* (2), 350−361.
(33) Samways, M. L.; Bruce Macdonald, H. E.; Taylor, R. D.; Essex, J. W. Water Networks in Complexes between Proteins and FDA-Approved Drugs. *J. Chem. Inf. Model.* **2023**, *63* (1), 387−396.
(34) Spyrakis, F.; Ahmed, M. H.; Bayden, A. S.; Cozzini, P.; Mozzarelli, A.; Kellogg, G. E. The Roles of Water in the Protein Matrix: A Largely Untapped Resource for Drug Discovery. *J. Med. Chem.* **2017**, *60* (16), 6781−6827.
(35) Clarke, C.; Woods, R. J.; Gluska, J.; Cooper, A.; Nutley, M. A.; Boons, G.-J. Involvement of Water in Carbohydrate−Protein Binding. *J. Am. Chem. Soc.* **2001**, *123* (49), 12238−12247.
(36) Pastor, M.; Cruciani, G.; Watson, K. A. A Strategy for the Incorporation of Water Molecules Present in a Ligand Binding Site into a Three-Dimensional Quantitative Structure−Activity Relationship Analysis. *J. Med. Chem.* **1997**, *40* (25), 4089−4102.
(37) Chen, J. M.; Xu, S. L.; Wawrzak, Z.; Basarab, G. S.; Jordan, D. B. Structure-Based Design of Potent Inhibitors of Scytalone Dehydratase: Displacement of a Water Molecule from the Active Site. *Biochemistry* **1998**, *37* (51), 17735−17744.
(38) Kadirvelraj, R.; Foley, B. L.; Dyekjær, J. D.; Woods, R. J. Involvement of Water in Carbohydrate−Protein Binding: Concanavalin A Revisited. *J. Am. Chem. Soc.* **2008**, *130* (50), 16933−16942.
(39) Matricon, P.; Suresh, R. R.; Gao, Z.-G.; Panel, N.; Jacobson, K. A.; Carlsson, J. Ligand design by targeting a binding site water. *Chem. Sci.* **2021**, *12* (3), 960−968.
(40) Bairagya, H. R.; Mukhopadhyay, B. P.; Bhattacharya, S. Role of the conserved water molecules in the binding of inhibitor to IMPDH-II (human): A study on the water mimic inhibitor design. *J. Mol. Struct.* **2009**, *908* (1), 31−39.
(41) Shaltiel, S.; Cox, S.; Taylor, S. S. Conserved water molecules contribute to the extensive network of interactions at the active site of protein kinase A. *Proc. Natl. Acad. Sci. U. S. A.* **1998**, *95* (2), 484−491.
(42) Biela, A.; Khayat, M.; Tan, H.; Kong, J.; Heine, A.; Hangauer, D.; Klebe, G. Impact of Ligand and Protein Desolvation on Ligand Binding to the S1 Pocket of Thrombin. *J. Mol. Biol.* **2012**, *418* (5), 350−366.
(43) Meine, R.; Becker, W.; Falke, H.; Preu, L.; Loaëc, N.; Meijer, L.; Kunick, C. Indole-3-Carbonitriles as DYRK1A Inhibitors by Fragment-Based Drug Design. *Molecules* **2018**, *23* (2), 64.
(44) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28* (1), 235−242.
(45) Wilcken, R.; Zimmermann, M. O.; Lange, A.; Joerger, A. C.; Boeckler, F. M. Principles and Applications of Halogen Bonding in Medicinal Chemistry and Chemical Biology. *J. Med. Chem.* **2013**, *56* (4), 1363−1388.
(46) *Python Programming Language*, version 3.9. https://www.python.org/.
(47) *The PyMOL Molecular Graphics System*, Version 2.5.5: Schrödinger, LLC. 2015.
(48) Yang, Y.; Fan, X.; Liu, Y.; Ye, D.; Liu, C.; Yang, H.; Su, Z.; Zhang, Y.; Liu, Y. Function and inhibition of DYRK1A: Emerging roles of treating multiple human diseases. *Biochem. Pharmacol.* **2023**, *212*, 115521.
(49) *Schrödinger Release 2021−1*; Schrödinger, LLC, New York, NY, 2021.
(50) Staroverov, V. N.; Scuseria, G. E.; Tao, J.; Perdew, J. P. Comparative assessment of a new nonempirical density functional: Molecules and hydrogen-bonded complexes. *J. Chem. Phys.* **2003**, *119* (23), 12129−12137.
(51) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *J. Chem. Phys.* **2010**, *132* (15), 154104.
(52) Schäfer, A.; Huber, C.; Ahlrichs, R. Fully optimized contracted Gaussian basis sets of triple zeta valence quality for atoms Li to Kr. *J. Chem. Phys.* **1994**, *100* (8), 5829−5835.
(53) *TURBOMOLE V7.4.1 2019, a development of University of Karlsruhe and Forschungszentrum Karlsruhe GmbH*, 1989−2007, TURBOMOLE GmbH, since 2007.
(54) Moller, C.; Plesset, M. S. Note on an Approximation Treatment for Many-Electron Systems. *Phys. Rev.* **1934**, *46* (7), 618−622.
(55) Metrangolo, P.; Murray, J. S.; Pilati, T.; Politzer, P.; Resnati, G.; Terraneo, G. Fluorine-Centered Halogen Bonding: A Factor in Recognition Phenomena and Reactivity. *Cryst. Growth Des.* **2011**, *11* (9), 4238−4246.
(56) *PDB - Small Molecule Statistics: Molecular Weight Distribution*. https://www.rcsb.org/stats/chemcomp/distribution-chem-comp-molecular-weight (accessed 2023).

(57) Shao, C.; Westbrook, J. D.; Lu, C.; Bhikadiya, C.; Peisach, E.; Young, J. Y.; Duarte, J. M.; Lowe, R.; Wang, S.; Rose, Y.; et al. Simplified quality assessment for small-molecule ligands in the Protein Data Bank. *Structure* **2022**, *30* (2), 252−262.e254.
(58) Merz, K. M., Jr. Using Quantum Mechanical Approaches to Study Biological Systems. *Acc. Chem. Res.* **2014**, *47* (9), 2804−2811.
(59) Shurki, A.; Warshel, A. Structure-Function Correlations of Proteins using MM, QM-MM, and Related Approaches: Methods, Concepts, Pitfalls, and Current Progress. In *Advances in Protein Chemistry*, Vol. *66*; Academic Press, 2003; pp 249−313.
(60) Dammann, M.; Stahlecker, J.; Zimmermann, M. O.; Klett, T.; Rotzinger, K.; Kramer, M.; Coles, M.; Stehle, T.; Boeckler, F. M. Screening of a Halogen-Enriched Fragment Library Leads to Unconventional Binding Modes. *J. Med. Chem.* **2022**, *65* (21), 14539−14552.
(61) Kabsch, W. XDS. *Acta Crystallogr. Sect. D. Biol. Crystallogr.* **2010**, *66* (2), 125−132.
(62) McCoy, A. J.; Grosse-Kunstleve, R. W.; Adams, P. D.; Winn, M. D.; Storoni, L. C.; Read, R. J. Phaser crystallographic software. *J. Appl. Crystallogr.* **2007**, *40* (4), 658−674.
(63) Winn, M. D.; Ballard, C. C.; Cowtan, K. D.; Dodson, E. J.; Emsley, P.; Evans, P. R.; Keegan, R. M.; Krissinel, E. B.; Leslie, A. G. W.; McCoy, A.; et al. Overview of the CCP4 suite and current developments. *Acta Crystallogr. Sect. D. Biol. Crystallogr.* **2011**, *67* (4), 235−242.
(64) Soundararajan, M.; Roos, A. K.; Savitsky, P.; Filippakopoulos, P.; Kettenbach, A. N.; Olsen, J. V.; Gerber, S. A.; Eswaran, J.; Knapp, S.; Elkins, J. M. Structures of Down Syndrome Kinases, DYRKs, Reveal Mechanisms of Kinase Activation and Substrate Recognition. *Structure* **2013**, *21* (6), 986−996.
(65) Emsley, P.; Lohkamp, B.; Scott, W. G.; Cowtan, K. Features and development of Coot. *Acta Crystallogr. Sect. D. Biol. Crystallogr.* **2010**, *66* (4), 486−501.
(66) Liebschner, D.; Afonine, P. V.; Baker, M. L.; Bunkoczi, G.; Chen, V. B.; Croll, T. I.; Hintze, B.; Hung, L.-W.; Jain, S.; McCoy, A. J.; et al. Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix. *Acta Crystallogr. Sect. D. Biol. Crystallogr.* **2019**, *75* (10), 861−877.
(67) Long, F.; Nicholls, R. A.; Emsley, P.; Grazulis, S.; Merkys, A.; Vaitkus, A.; Murshudov, G. N. AceDRG: a stereochemical description generator for ligands. *Acta Crystallogr. Sect. D. Biol. Crystallogr.* **2017**, *73* (2), 112−122.
(68) Weigend, F.; Häser, M. RI-MP2: first derivatives and global consistency. *Theor. Chem. Acc.* **1997**, *97* (1), 331−340.
(69) Weigend, F.; Häser, M.; Patzelt, H.; Ahlrichs, R. RI-MP2: optimized auxiliary basis sets and demonstration of efficiency. *Chem. Phys. Lett.* **1998**, *294* (1), 143−152.
(70) Hättig, C. Geometry optimizations with the coupled-cluster model CC2 using the resolution-of-the-identity approximation. *J. Chem. Phys.* **2003**, *118* (17), 7751−7761.
(71) Hoffmann, R. An Extended Hückel Theory. I. Hydrocarbons. *J. Chem. Phys.* **2004**, *39* (6), 1397−1412.
(72) Weigend, F.; Ahlrichs, R. Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy. *Phys. Chem. Chem. Phys.* **2005**, *7* (18), 3297−3305.
(73) Hättig, C.; Hellweg, A.; Köhn, A. Distributed memory parallel implementation of energies and gradients for second-order Møller−Plesset perturbation theory with the resolution-of-the-identity approximation. *Phys. Chem. Chem. Phys.* **2006**, *8* (10), 1159−1169.
(74) Häser, M.; Ahlrichs, R. Improvements on the direct SCF method. *J. Comput. Chem.* **1989**, *10* (1), 104−111.
(75) Feyereisen, M.; Fitzgerald, G.; Komornicki, A. Use of approximate integrals in ab initio theory. An application in MP2 energy calculations. *Chem. Phys. Lett.* **1993**, *208* (5), 359−363.

https://doi.org/10.1021/acs.jctc.4c00834
J. Chem. Theory Comput. 2024, 20, 10716−10730

Appendix B: Publication 2

Comparison of QM Methods for the Evaluation of Halogen-π Interactions for Large-Scale Data Generation

Marc U. Engelhardt, Markus O. Zimmermann, Finn Mier, Frank M. Boeckler;

Journal of Chemical Theory and Computation, 2025

DOI: 10.1021/acs.jctc.5c00456

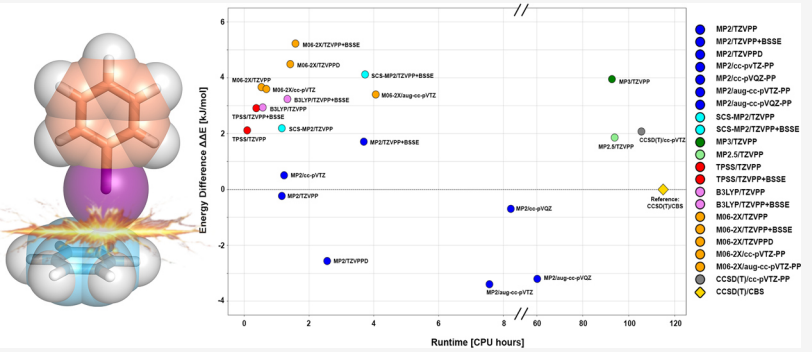
JCTC Journal of Chemical Theory and Computation

pubs.acs.org/JCTC Article

Comparison of QM Methods for the Evaluation of Halogen−π Interactions for Large-Scale Data Generation

Marc U. Engelhardt, Markus O. Zimmermann, Finn Mier, and Frank M. Boeckler*

Cite This: J. Chem. Theory Comput. 2025, 21, 6174−6183 Read Online



ABSTRACT: Halogen−π interactions play a pivotal role in molecular recognition processes, drug design, and therapeutic strategies, providing unique opportunities for enhancing and fine-tuning the binding affinity and specificity of pharmaceutical agents. The present study systematically benchmarks various combinations of quantum mechanical (QM) methods and basis sets to characterize halogen−π interactions in model systems. We evaluate both density functional theory (DFT) methods and wave function-based post-HF methods in terms of accuracy to reference calculations at the CCSD(T)/CBS level of theory and runtime efficiency. By balancing these crucial aspects, we aim to identify an optimal configuration suitable for high-throughput applications. Our results indicate that MP2 using the reasonably large TZVPP basis set is in excellent agreement with reference calculations, striking a balance between accuracy and computational efficiency. This allows us to generate large, reliable data sets, which will serve as a basis to develop and train machine-learning models capable of accurately capturing the strength of halogen−π interactions, thereby providing a robust data-driven foundation for medicinal chemistry analysis.

INTRODUCTION

Noncovalent interactions play a fundamental role in biological systems and molecular recognition processes, serving as the keystone for understanding biomolecular function, drug binding, and protein−ligand interactions.[1−5] Among these, halogen bonding (XB) has emerged as a unique and versatile interaction, characterized by the directional attraction between an electrophilic region on a halogen atom (σ-hole), typically chlorine, bromine, or iodine, and a nucleophilic partner.[6−12] These interactions have proven particularly useful in medicinal chemistry and drug design, where they not only enhance the binding affinity and specificity of ligands and stability of protein−ligand complexes[13−21] but also can contribute to ligands engaging in unconventional binding modes.[22−24]

Various nucleophilic moieties that form noncovalent interactions with a halogen atom in the protein binding site have been systematically investigated.[25−27] Although XB acceptors such as backbone carbonyls and the π-surface of the peptide bond,[28,29] the sulfur atom in methionine,[30] the nitrogen atoms in histidine,[31] carboxylate (aspartate/glutamate) and carboxamide (asparagine/glutamine) moieties,[32] and oxygen atoms of water molecules[33,34] have already been studied extensively, the importance of addressing π-systems (aromatic side chains of tyrosine, phenylalanine, histidine, and tryptophan) as XB acceptors in protein−ligand interactions has only been highlighted and systematic approaches are still underrepresented.[35−38]

To date, the nature of halogen bonding is still a controversial subject, and thus, theoretical calculations are of tremendous importance.[39−43] In 2012, Rezac et al.[44] analyzed and benchmarked calculations of various noncovalent interactions of halogenated molecules using different quantum mechanical (QM) methods and basis sets and generated a small benchmark set of interaction geometries. With respect to halogen−π interactions, halomethane or a tuned variant,

Received: March 20, 2025
Revised: May 6, 2025
Accepted: May 27, 2025
Published: June 9, 2025

https://doi.org/10.1021/acs.jctc.5c00456
J. Chem. Theory Comput. 2025, 21, 6174−6183

trifluorohalomethane, in complex with a benzene molecule, was used. In 2020, Zhu et al.[45] published a perspective on the application of QM methods to evaluate halogen bonding, where they further investigated halogen interactions in general, including aromatic systems as acceptors. Despite their initial and pioneering work, in-depth analysis on a larger scale, especially for halogen bonding donors and acceptors with relevance to the drug discovery process, is still limited. Accurate modeling of these interactions is essential for capturing their energetic contributions to protein−ligand binding and their application in guiding structure-based drug design. Wallnoefer et al.[38] investigated interactions of chlorobenzene and bromobenzene, addressing a *p*-cresol system, and provided an initial comparison of QM methods and basis sets for such systems.

QM methods differ in how they treat electronic interactions, electron correlation, and exchange effects, impacting their accuracy and computational cost.[46] Balancing these factors is crucial, especially when dealing with complex biological systems. Among several methods, the coupled cluster method CCSD(T) is widely used as the "gold standard" because it is the most accurate, nonempirical method applicable to reasonably large systems of practical interest.[47] Møller-Plesset perturbation theory (MPn)[48] methods are post-Hartree−Fock approaches that explicitly account for electron correlation by systematically improving the wave function obtained from the Hartree−Fock calculation using *n*th order perturbation theory. MP2[49] (second-order Møller-Plesset) is widely used for its balance between accuracy and computational cost. MP3[50,51] extends this approach by including third-order terms, offering improved accuracy but at an exponentially higher computational cost. MP2.5 is a pragmatic compromise that averages MP2 and MP3 energies, often providing results closer to CCSD(T) with reduced computational requirements in comparison to coupled cluster calculations.[52] This approach exploits the systematic error compensation between MP2's tendency to underestimate and MP3's tendency to overestimate in some systems. Spin-Component-Scaled MP2 (SCS-MP2[53,54]) refines MP2 by applying different scaling factors to the parallel spin and opposite spin electron correlation components. This adjustment improves the accuracy by reducing the tendency to overestimate electron correlation effects by MP2.

In contrast, density functional theory methods (DFT) approximate the electron correlation through exchange-correlation functionals based on electron density rather than the wave function.[55] TPSS[56] is a GGA[57] (generalized gradient approximation) functional that includes a kinetic energy density term, improving the accuracy for weak interactions and transition metal chemistry. B3LYP[58,59] and M06−2X[60] are popular hybrid GGA functionals, mixing parts of the exact Hartree−Fock exchange with DFT exchange-correlation functionals, to balance computational cost and accuracy. Accuracy of the DFT methods can be enhanced by adding Grimme's D3 dispersion correction for noncovalent interactions.[61]

Besides the choice of an appropriate QM method, a reasonable choice of the basis set for the calculation is certainly important. Commonly used basis sets include the triple-ζ valence with polarization (TZVPP[62]) or an enhanced variant with an additional diffuse function (TZVPPD), the correlation-consistent polarized valence X-ζ[63] (cc-pVXZ, where X = *D*, *T*, *Q*, etc.), and its extended counterpart, aug-cc-pVXZ.[64] The TZVPP basis set, widely used in density functional theory (DFT) and post-Hartree−Fock methods, provides a robust trade-off between computational efficiency and accuracy by including multiple polarization functions. The cc-pVXZ family of basis sets, on the other hand, is designed to improve electron correlation effects with increasing cardinality X. The extended versions (aug-cc-pVXZ) also include diffuse functions, which further enhance contributions of dispersion effects for noncovalent interactions. The suffix "-PP" indicates an additional pseudopotential for certain higher-order atoms, such as iodine. Although such calculations are practically impossible, the most accurate result would be achieved by CCSD(T) calculations using a complete basis set (CBS). To address this issue, calculations using smaller basis sets, typically correlation-consistent basis sets, can be extrapolated to the complete basis set limit.[65]

In this study, we focus on the systematic investigation of halogen−π interactions using high-level quantum mechanical (QM) methods, including post-Hartree−Fock and DFT, in combination with commonly used basis sets. We aim to select a proper method with a reasonable balance between speed and accuracy, in order to apply this method for generating big data sets of several million halogen−π interaction geometries. Based on this big data, we will strive to derive models from machine learning approaches that enable us to predict the interaction energy for a given geometry almost instantly without the need for calculations at the QM level.

A systematic grid of iodobenzene as a ligand model system in complex with benzene as a halogen bond acceptor is generated, and single-point calculations are carried out. Our focus clearly is on iodobenzene, as a complementary halogen bond donor to previous studies,[38] but also, particularly, because it provides the strongest halogen bonds in comparison to bromobenzene or chlorobenzene. To ensure comparability and applicability among the halobenzenes, single-point calculations for combinations of methods and basis sets of particular importance were conducted for chlorobenzene and bromobenzene as well. Calculations on the CCSD(T) level of theory extrapolated to the complete basis set limit serve as a reference. We report energy differences and computational costs (CPU runtime) to evaluate the most suitable method and basis set combination for representing halogen−π interactions. From previous studies,[27,28,30−32] we have often observed good applicability of calculations on an MP2/TZVPP level of theory. Thus, we were interested to see whether this experience could be transferred to noncovalent interactions with π-systems.

RESULTS AND DISCUSSION

Comparison of QM Methods and Basis Sets. In comparison to interactions involving single atom acceptors, such as oxygen or nitrogen, halogen−π interactions in haloaryl systems (where the halogen atom is directly attached to an aromatic ring) exhibit greater structural diversity. This is based on the strong increase of possible interaction geometries, correlating to the larger π-surface for forming attractive interactions. This difference arises from the extended π-plane of the aromatic systems, which provides a delocalized electron cloud capable of interacting with the σ-hole of the halogen atom in multiple geometric orientations. In contrast, the localized lone pair electrons of oxygen or nitrogen constrain the halogen bond donors to fewer but more specific orientation geometries, showing less flexibility than the π-systems. Still, the extended π-surface is also a "double-edged

https://doi.org/10.1021/acs.jctc.5c00456
J. Chem. Theory Comput. 2025, 21, 6174−6183

Table 1. Mean Energy Difference, $\Delta\Delta E$, Mean Absolute Energy Differences ($|\Delta\Delta E|$), and RMSD of Different QM Methods and the Used Basis Sets[b]

method	basis set	mean $\Delta\Delta E$ to CCSD(T)/CBS	abs. mean $\|\Delta\Delta E\|$ to CCSD(T)/CBS	RMSD to CCSD(T)/CBS	mean runtime (CPUh)
MP2	TZVPP	−0.23	0.64	0.91	1.16
	TZVPP+BSSE	1.71	1.71	2.43	3.69
	TZVPPD	−2.56	2.57	3.51	2.56
	cc-pVTZ-PP	0.51	0.59	0.73	1.23
	cc-pVQZ-PP	−0.69	0.96	1.66	8.23
	aug-cc-pVTZ-PP	−3.39	3.39	4.44	7.57
	aug-cc-pVQZ-PP	−3.20	3.20	4.39	60.14
SCS-MP2	TZVPP	2.19	2.19	3.23	1.16
	TZVPP+BSSE	4.12	4.11	6.18	3.73
MP3	TZVPP	3.95	3.95	6.22	92.68
MP2.5	TZVPP	1.86	1.85	2.83	93.84
TPSS (-D3)	TZVPP	2.12	2.11	4.25	0.09
	TZVPP+BSSE	2.91	2.91	4.81	0.37
B3LYP (-D3)	TZVPP	2.94	2.93	6.25	0.57
	TZVPP+BSSE	3.24	3.23	6.53	1.33
M06−2X (-D3)	TZVPP	3.67	3.66	5.10	0.53
	TZVPP+BSSE	5.23	5.21	6.57	1.58
	TZVPPD	4.49	4.51	11.44	1.42
	cc-pVTZ-PP	3.60	3.59	5.10	0.68
	aug-cc-pVTZ-PP	3.40	3.40	4.86	4.06
CCSD(T)	cc-pVTZ-PP	2.08	2.08	3.46	105.51
CCSD(T)	CBS	[a]	[a]	[a]	114.97

[a]No values reported, since CCSD(T)/CBS is the reference for all other methods. [b]Energy values are given as difference between corresponding method values and the reference calculations of CCSD(T)/CBS in kJ/mol. Runtime is given in CPU hours.

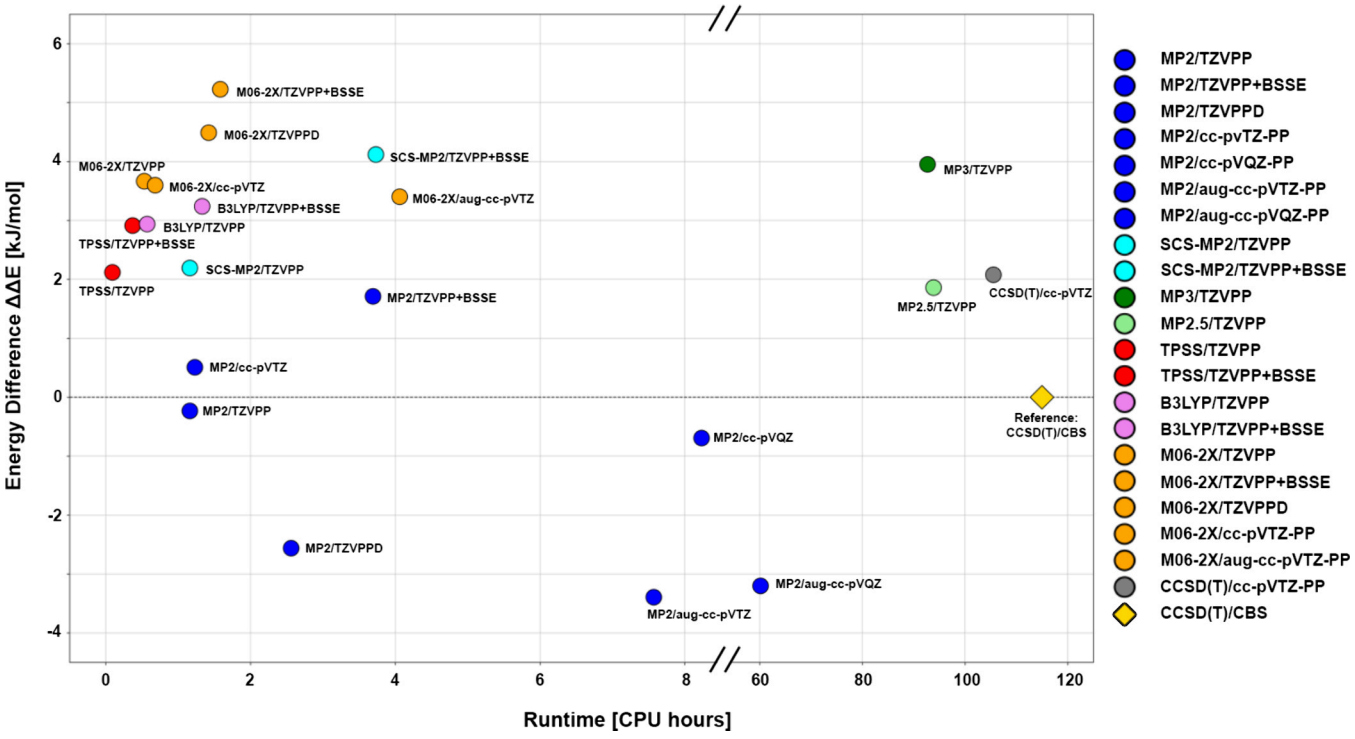


Figure 1. Mean energy difference $\Delta\Delta E$ (in kJ/mol) to the reference CCSD(T)/CBS and runtime (in CPU hours) of the evaluated methods and basis sets of iodine. The dashed horizontal line at $\Delta\Delta E$ = 0 kJ/mol indicates the CCSD(T)/CBS reference level (golden diamond), while the *x*-axis break indicates the large jump in computational cost for higher-level methods. Each color corresponds to a different level of theory and basis set treatment as shown in the legend. The figure was prepared by using custom Python scripts and the *matplotlib* library.

sword", as it increases the risk for the formation of secondary interactions such as $\pi\cdots\pi$ or C−H$\cdots\pi$.

Researchers have reported different results and opinions regarding the suitability of different QM methods for halogen bonding. However, our group's previous investigations at the

https://doi.org/10.1021/acs.jctc.5c00456
J. Chem. Theory Comput. 2025, 21, 6174−6183

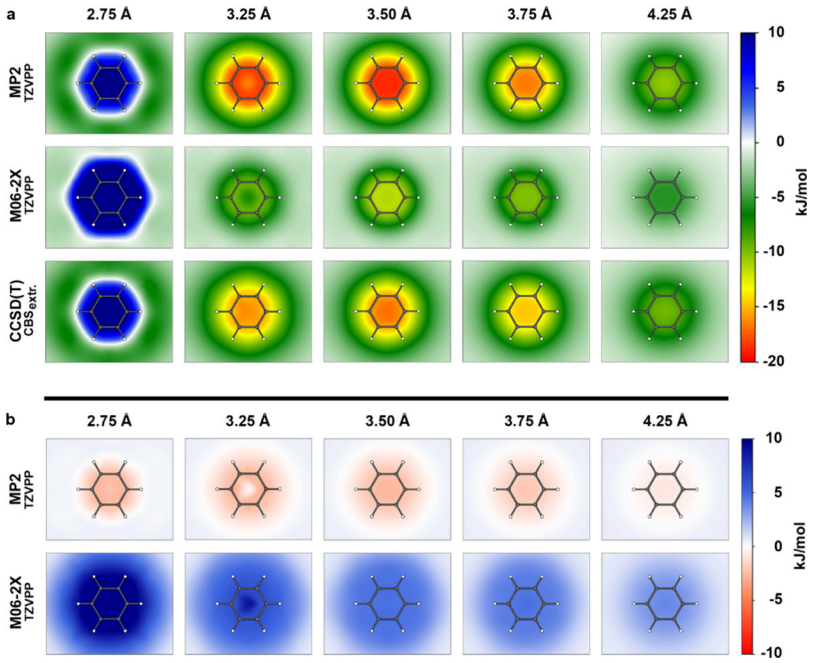


Figure 2. Adduct formation energy surfaces of MP2 and M06−2X with the TZVPP basis set, as well as the surfaces of the reference CCSD(T)/CBS. Surfaces represent the halogen−π interaction energies ΔE of iodobenzene in complex with the targeted benzene at distances of $d_{I\cdots\pi\text{-plane}}$ = [2.75, 3.25, 3.50, 3.75, 4.25 Å]. The iodobenzene is oriented perpendicular to the π-plane. Data points of the surface are interpolated and colored according to the given energy scale. (a) Surfaces of adduct formation energies ΔE. Positive energies and negative energies are capped to 10 and −20 kJ/mol, respectively, for better visibility. (b) Surfaces of the difference between adduct formation energies of MP2 and M06−2X and the reference CCSD(T)/CBS (calculated as $\Delta\Delta E = \Delta E_{method} - \Delta E_{CCSD(T)/CBS}$). Positive and negative differences were capped to 10 and −10 kJ/mol. Figures were prepared by using custom Python scripts and the *matplotlib* library.

MP2 level, using a triple-ζ (def2-TZVPP) basis set on different halogen bond acceptors, yielded accurate adduct formation energies while maintaining a feasible computational time. To ensure that this level of calculation is also suitable for halogen−π interactions, we conducted benchmark calculations of different combinations of QM methods and basis sets. Adduct formation energies were compared to high-level reference calculations on a CCSD(T) level of theory, extrapolated to the basis set limit using the approach proposed by Halkier et al.[65] It should be noted that we will use the short-term "CCSD(T)/CBS" subsequently as always referring to this complete basis set extrapolation approach. At this level of theory, only a smaller subset (∼30% of all geometries) was used due to the extraordinarily high computational cost. Several applied methods were counterpoise corrected (BSSE correction) with the procedure of Boys and Bernardi.[66] However, it has to be noted that the effectiveness of BSSE correction remains a controversial subject in the literature.[67]

Table 1 shows the mean energy deviations ($\Delta\Delta E$), the mean absolute energy deviations ($|\Delta\Delta E|$), and the root-mean-square deviations (RMSD) in kJ/mol from the reference CCSD(T)/CBS, together with the corresponding computational cost in CPU hours. A detailed table incorporating the individual data points and energies of all methods is provided as a separate Excel file, which can be found as part of the Supporting Information. Mean differences, calculated as $\Delta\Delta E = \Delta E_{method} - \Delta E_{CCSD(T)/CBS}$, where ΔE denotes the adduct formation energy, indicate an overall deviation between the methods and their tendency to over- or underestimate energies. However, these values can be misleading, as large positive and negative values can level each other out. Therefore, we additionally report the mean absolute energy deviation and the RMSD. RMSD values can give further insight into the magnitude of the large difference compared to $|\Delta\Delta E|$. The runtime is averaged over single-points of different distances and across the calculated grid. It is obvious that the CCSD(T) reference calculations require by far the highest resources, with about 115 h per single-point. The majority of the computational costs for extrapolation are caused by the CCSD(T)/cc-pVTZ-PP calculation (105.51 h).

Among the evaluated methods, MP2 using the TZVPP basis set stands out with an excellent balance between accuracy and computational cost (Figure 1). The method shows mean deviations and absolute mean deviations of $\Delta\Delta E$ = −0.23 kJ/mol and $|\Delta\Delta E|$ = 0.64 kJ/mol, respectively. This level of accuracy is achieved with a notably low computational cost of about 1.16 CPU hours on average per single-point calculation. In contrast, the BSSE-corrected version of MP2/TZVPP results in larger deviations ($\Delta\Delta E = |\Delta\Delta E|$ = 1.71 kJ/mol, RMSD = 2.43 kJ/mol) and requires significantly more time with 2.56 CPU h, due to the calculations of ghost molecules to eliminate overlapping terms. Calculations using the diffuse function enhanced TZVPPD basis set, unfortunately, show the worst results of the triple-ζ variants with $\Delta\Delta E$ = −2.56 kJ/mol, $|\Delta\Delta E|$ = 2.57 kJ/mol, and RMSD = 3.51 kJ/mol. The correlation-consistent basis sets cc-pVTZ-PP and cc-pVQZ-PP also show very good results with $\Delta\Delta E$ = 0.59 kJ/mol (RMSD = 0.73 kJ/mol) and $\Delta\Delta E$ = 0.96 kJ/mol (RMSD = 1.66 kJ/mol), respectively. However, looking at the runtime, cc-pVTZ-PP with 1.23 CPU hours may still compete with TZVPP, while the larger cc-pVQZ-PP basis set with 8.23 CPU hours on average seems neither efficient nor most effective. Furthermore, augmented basis sets aug-cc-pVTZ-PP and aug-cc-pVQZ-PP both perform quite similarly for all energy results,

https://doi.org/10.1021/acs.jctc.5c00456
J. Chem. Theory Comput. 2025, 21, 6174−6183

Table 2. Mean Energy Difference, $\Delta\Delta E$, Mean Absolute Energy Differences ($|\Delta\Delta E|$), and RMSD of Different QM Methods for Chlorine and Bromine Interactions[b]

method	basis set	mean $\Delta\Delta E$ to CCSD(T)/CBS	abs mean $\Delta\Delta E$ to CCSD(T)/CBS	RMSD to CCSD(T)/CBS	mean runtime (CPUh)
Chlorine					
MP2	TZVPP	−0.68	0.68	1.06	1.01
	TZVPP+BSSE	0.35	0.35	0.51	3.23
M06−2X (-D3)	TZVPP	0.84	0.85	1.20	0.49
CCSD(T)	CBS	[a]	[a]	[a]	79.28
Bromine					
MP2	TZVPP	−0.79	0.79	1.30	1.15
	TZVPP+BSSE	0.32	0.33	0.50	3.56
M06−2X (-D3)	TZVPP	1.35	1.35	1.86	0.57
CCSD(T)	CBS	[a]	[a]	[a]	105.37

[a]No values reported, since CCSD(T)/CBS is the reference for all other methods. [b]Energy values are given as difference between corresponding method values and the reference calculations of CCSD(T)/CBS in kJ/mol. Runtime is given in CPU hours.

but the deviations here are rather large and with much higher runtimes of 7.57 and even over 60 CPU hours, respectively.

As an alternative post-Hartree−Fock method, MP3 should give relatively accurate predictions with slightly underestimated energy values.[68] With an energy difference of $|\Delta\Delta E|$ = 3.95 kJ/mol (RMSD = 6.22 kJ/mol) and the vast computational effort requiring almost 93 CPU hours, however, it is less practical for routine calculations. Although MP2.5, as the arithmetic mean of MP2 and MP3, should compensate for over- and underestimations of both methods respectively, it shows higher deviations ($|\Delta\Delta E|$ = 1.85 kJ/mol, RMSD = 2.83 kJ/mol) with the runtime obviously dominated by the MP3 calculations.[69] However, calculation of MP2.5 energies is, of course, cheap if both MP2 and MP3 calculations are conducted.

Computationally less demanding DFT methods, including B3LYP(-D3) and TPSS(-D3), generally show larger absolute deviations in this comparison ($|\Delta\Delta E|$ = 2.9 and 2.1 kJ/mol, respectively). Incorporating BSSE correction even increases the difference in energy values. In previous studies, the widely used M06−2X functional showed very accurate results, especially for weak interactions with dispersion contribution (including halogen bonding).[70,71] Therefore, the comparison between M06−2X and MP2 across different basis sets is of high interest. Although computationally efficient, with runtimes ranging from 0.53 to 4 h depending on the basis set, M06−2X generally shows larger deviations from the CCSD(T)/CBS reference than MP2. For example, M06−2X/TZVPP shows an absolute deviation of 3.66 kJ/mol (RMSD = 5.1 kJ/mol), and the BSSE-corrected version further increases this deviation to 5.21 kJ/mol (RMSD = 6.57 kJ/mol). This indicates that M06−2X may be suitable for highlighting tendencies but lacks the necessary quantitative accuracy for this application. M06−2X/TZVPPD shows trends similar to those of MP2/TZVPPD in terms of increasing the difference even further. M06−2X/TZVPPD even shows the highest RMSD among all of the tested methods. Using cc-pVTZ-PP and aug-cc-pVTZ-PP yields similarly inaccurate results as MP2 using the same basis sets.

"2D energy surface plots" of the actual adduct formation energies ΔE were generated for each of the five different distances individually to highlight attractive and repulsive areas. This means that the surface in plane with the aromatic ring system of benzene is colored at the position of the halogen atom above this plane based on the ΔE value for this halogen−π interaction, followed by interpolating between these energies. Furthermore, we generated 2D surface plots of $\Delta\Delta E$, as well, showing the deviation from the reference calculations of $\Delta\Delta E$ in a similar fashion. For simplicity, here we only compare MP2/TZVPP and M06−2X/TZVPP. Surface plots of ΔE and $\Delta\Delta E$ of the remaining methods can be found in the Supporting Information Figures S1 and S2. Figure 2a shows the adduct formation energy surface plots of MP2/TZVPP and M06−2X/TZVPP for all investigated distances (2.75 3.25, 3.50, 3.75, and 4.25 Å) individually as well as the surfaces of the reference CCSD(T)/CBS energies. For MP2/TZVPP, adduct formation energies ΔE range from −18.88 kJ/mol as the most favorable interaction to 31.19 kJ/mol as highly repulsive. Use of M06−2X/TZVPP provides ranges from −11.59 to 53.55 kJ/mol, while "gold standard" CCSD(T)/CBS yields a range of adduct formation energies from −16.17 to 33.43 kJ/mol. For better visibility, positive energy values were capped at 10 kJ/mol. For d = 2.75 Å, mainly repulsive or only minimal attractive interactions can be observed. Increasing the distance rapidly shifts the interaction from repulsive to attractive. Most favorable interactions with minimum energy values can be observed at d = 3.5 Å for both methods. Figure 2b shows the energy surface based on the adduct formation energy difference $\Delta\Delta E$ between the two methods and CCSD(T)/CBS for all distances. Positive and negative values were capped at 10 and −10 kJ/mol, respectively. Original values are provided in spreadsheet format (xlsx) in the Supporting Information. MP2/TZVPP shows very low differences to the reference calculation, ranging from $\Delta\Delta E$ = 0.59 to −2.9 kJ/mol, while for M06−2X/TZVPP, deviations from the reference energies range from $\Delta\Delta E$ = 0.58 to 20.11 kJ/mol. It can be argued that precise predictions of highly repulsive energies are less relevant for drug discovery purposes as long as the strong repulsion is recognized and the area of the transition between attractive and repulsive interactions is not strongly altered. Thus, for benchmarking purposes, we keep them in the data set.

In summary, we find MP2 to be in very good agreement with the "gold standard" CCSD(T)/CBS across multiple basis sets, with absolute deviations as low as 0.64 kJ/mol for MP2/TZVPP and 0.59 kJ/mol for MP2/cc-pVTZ-PP. RMSD values also show minimal differences to the reference with 0.91 and 0.73 kJ/mol, respectively, with cc-pVTZ-PP performing slightly better. Thus, MP2, using either TZVPP or cc-pVTZ-PP, having almost identical levels of accuracy while maintaining feasible computational effort, seems an excellent

https://doi.org/10.1021/acs.jctc.5c00456
J. Chem. Theory Comput. 2025, 21, 6174−6183

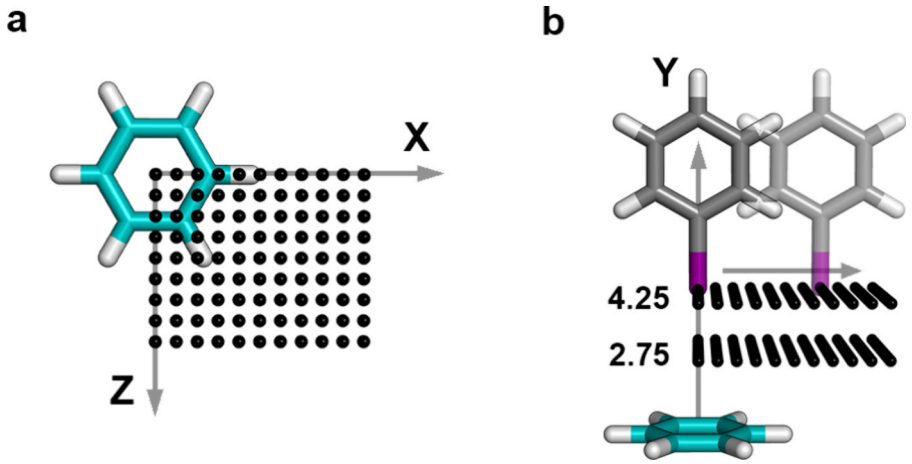


Figure 3. Overview of the interaction geometry generation on a regular grid. (a) Grid points on the *XZ*-plane were generated with dimensions $X_{translation}$ = [0.0−5.0 Å], $Z_{translation}$ = [0.0−4.0 Å] in steps of 0.5 Å. (b) Grid points were generated for five different distances $d_{I\cdots\pi\text{-plane}}$ between the halogen atom and the benzene plane, $d_{I\cdots\pi\text{-plane}}$ = [2.75 3.25, 3.5, 3.75, and 4.25 Å]. In this distance definition ($d_{I\cdots\pi\text{-plane}}$) the respective point on the π-plane is individually determined by the normalization to the plane through the iodine atom. Figures were prepared with PyMOL.[82]

choice. Given the overall results and the shorter runtime of MP2/TZVPP (1.16 h) compared to MP2/cc-pVTZ-PP (1.23 h), which could amount to saving several million CPU hours for large data sets, it was emphasized that our previous choice of MP2/TZVPP is a quite reasonable approach.

Additional Calculations with Chlorobenzene and Bromobenzene. Since interactions of chlorobenzene and bromobenzene with π-systems have been studied previously,[38,44,45] the focus of this study mainly lies on iodine interactions. Iodine has emerged as a particularly interesting element in medicinal chemistry because its large σ-hole enables the formation of exceptionally strong and highly directional halogen bonds, which medicinal chemists can exploit to modulate the binding affinity, target selectivity, and physicochemical properties of drug candidates. Although their halogen bonding ability is weaker than iodine's, traditionally bromine and chlorine remain more prevalent due to their milder steric impact and favorable synthetic versatility. In computational studies, iodine is typically modeled with a relativistic effective-core potential to account for its heavy-atom inner electrons. When benchmarked, conclusions gained for iodine can be confidently extended to its lighter halogen colleagues, bromine and chlorine, whose smaller relativistic contributions arise from the same underlying interactions. To ensure comparability and applicability, single-point calculations of chlorobenzene and bromobenzene in complex with benzene were performed at the MP2 and M06−2X levels of theory using the basis set TZVPP, as well as CCSD(T)/CBS extrapolation as a reference. The same set of 150 geometries was used for this comparison, applying a proper shift of the halobenzene scaffold to keep the halogen-π distance always identical to the iodine data set. Table 2 shows the results of both chlorine and bromine interactions. Similar to iodine, we report mean energy deviations ($\Delta\Delta E$), mean absolute energy deviations ($|\Delta\Delta E|$), and root-mean-square deviations (RMSD) in kJ/mol from the reference CCSD(T)/CBS, together with the corresponding computational cost in CPU hours. A detailed table incorporating the individual data points and energies of all methods for both chlorine and bromine results can be found in the Supporting Information. It can be concluded that chlorine and bromine interactions behave similarly to those of iodine. Using MP2 with TZVPP yields comparably good results (agreement with reference calculations) for chlorine and bromine as for iodine, while maintaining very low computational costs of around 1−1.3 CPU hours. Interestingly, however, the application of the counterpoise correction differs for chlorine and bromine interactions and shows even better results with lower energy differences from the reference. Unfortunately, MP2/TZVPP+BSSE still shows runtimes of more than 3-fold compared to MP2/TZVPP and thus appears less applicable to the calculation of very large data sets. Looking at the results of M06−2X(-D3) calculations, trends similar to those of iodine can be derived. While showing rather low computational costs, the energy differences are doubled compared to MP2 calculations. Using the same visualization strategy as in Figure 2, we generated individual "2D energy surface plots" and energy difference plots for each method at every examined distance (2.75, 3.25, 3.50, 3.75, and 4.25 Å) which can be found in the Supporting Information (Figure S3 for chlorine and Figure S4 for bromine).

CONCLUSIONS

In this work, we investigated the potential of different QM methods to correctly assess halogen−π interactions with a focus on iodine. Adduct formation energy differences, $\Delta\Delta E$, between QM methods and reference calculations using CCSD(T)/CBS, as well as the average runtime of single-point calculations, were reported. Results show that MP2 with the reasonably large basis set TZVPP is an excellent choice and is in very good agreement with reference calculations while maintaining feasible computational demands. With this study, we aim to provide a solid basis for characterizing halogen−π interactions in ab initio approaches and beyond. Similar to our previous experience,[27,28,30−32] good performance of MP2/TZVPP appears to be transferable onto the evaluation of iodine−π systems. We were able to demonstrate that MP2/TZVPP remains a very good choice for chlorine and bromine interactions with the π-surface of benzene as well. It is interesting to note that applying a counterpoise correction enhances the accuracy of MP2/TZVPP for chlorine and bromine. However, the still very high computational demands make this method quite impractical for large data applications. Based on a good balance of accuracy and speed for MP2/TZVPP, this method will be employed to generate large data

https://doi.org/10.1021/acs.jctc.5c00456
J. Chem. Theory Comput. 2025, 21, 6174−6183

sets as a source for machine learning of this pharmaceutically interesting interaction. With such a QM-AI approach, high-accuracy interaction energies could become available on the millisecond scale.

COMPUTATIONAL METHODS

Structure Optimization. Geometry optimizations of the individual ligand model system (iodobenzene, chlorobenzene, and bromobenzene) and the amino acid model system of phenylalanine (benzene) were done at the MP2 level of theory using TURBOMOLE 7.7.1[72] with a triple-ζ basis set (def2-TZVPP). Calculations were performed in combination with the resolution of identity (RI) technique and the frozen core approximation. Frozen core orbitals were defined using default settings, where orbitals with energies below −3.0 au are considered core orbitals. SCF convergence criterion was increased to 10^{-8} hartree. Relativistic effects for iodine were considered by an effective core potential (ECP).[73−81]

Generation of Interaction Geometries. Interaction geometries of iodobenzene in complex with benzene were generated. Iodobenzenes were placed on a regular grid using *X*- and *Z*-translations (Figure 3a) for five different distances along the *Y*-axis (Figure 3b). Following previous approaches, an optimal σ-hole angle of $\alpha_{C-I\cdots\pi\text{-plane}} = 180°$ was used. In this angle definition ($\alpha_{C-I\cdots\pi\text{-plane}}$), the respective point on the π-plane is individually determined by the normal to the plane through the iodine atom. Due to the symmetric nature of benzene, only one quadrant of the grid was considered. With this procedure, a total of 495 interaction geometries were generated to carry out a single-point calculation. For the comparison to CCSD(T)/CBS reference calculations, the same smaller subsets (∼30% of all geometries) were used for all halobenzenes.

QM Methods, Basis Sets, and Adduct Formation Energies. An overview of the different methods and basis set combinations can be seen in Table 1. All single-point calculations were carried out using TURBOMOLE 7.7.1 on the JUSTUS2−bwHPC Cluster,[83] where a standard node has a 2 × Intel Xeon E6252 Gold (Cascade Lake) CPU (2.1 GHz base, 3.7 GHz max. accelerated) with 192GB or 384GB memory. Calculations were done in combination with the resolution of identity (RI) technique and the frozen core approximation, if applicable. Frozen core orbitals were defined using default settings, where orbitals with energies below −3.0 au are considered core orbitals. SCF convergence criterion was increased to 10^{-8} hartree. Relativistic effects for iodine were considered by an effective core potential (ECP). Methods of choice comprise MP2, MP3, SCS-MP2, B3LYP, M06−2X, and TPSS. For selected methods and basis set combinations (see Table 1), energy values were counterpoise corrected using the procedure of Boys and Bernardi to eliminate basis set superposition errors (BSSEs). Basis sets included in the study were the triple-ζ basis set def2-TZVPP and the diffuse function enhanced variant def2-TZVPPD. Further, the correlation consistent basis sets cc-pVNZ-PP and the augmented basis sets aug-cc-pVNZ-PP ($N = T, Q$) were used with an additional pseudo potential for iodine (denoted by the "-PP" suffix). The DFT functionals TPSS, B3LYP, and M06−2X were augmented with an empirical dispersion correction as proposed by Grimme et al.,[61] which is indicated by adding "(-D3)" to the name. For the previously investigated chlorobenzene and bromobenzene, a small set of methods and basis sets was applied to provide the possibility of comparing our data for iodine to both less heavy halogens.

As a reference, single-point calculations at the complete basis set limit approximation were carried out using an extrapolation scheme proposed by Halkier et al.[65] Higher-order correlation energy was calculated using the following equation:

$$\Delta E_{CBS}^{CCSD(T)} = \Delta E_{CBS}^{MP2} + (\Delta E^{CCSD(T)} - \Delta E^{MP2})_{cc-pVTZ-PP} \quad (1)$$

This is due to the assumption that the difference between the CCSD(T) and MP2 interaction energies depends only slightly on the basis set and can therefore be estimated using a small or medium basis set, such as cc-pVTZ-PP. ΔE_{CBS}^{MP2} represents the energy at the complete basis set limit and can be determined as follows:

$$\Delta E_{CBS}^{MP2} = \frac{\Delta E_X^{MP2} X^3 - \Delta E_Y^{MP2} Y^3}{X^3 - Y^3} \quad (2)$$

where *X* and *Y* denote the cardinal numbers of the cc-pVTZ-PP and cc-pVQZ-PP basis set ($T = 3$ and $Q = 4$). Adduct formation energies were calculated as

$$\Delta E = (E_{complex} - (E_{halobenzene} + E_{benzene})) \quad (3)$$

and reported as kJ/mol.

ASSOCIATED CONTENT

Data Availability Statement

PyMOL is an open-source software maintained and distributed by Schrödinger. There is an open-source version of PyMOL available at: https://github.com/schrodinger/pymol-open-source. Python and all of its' packages is an open-source programming language available and downloadable from https://www.python.org/. Detailed results of the QM method comparison and an overview of the individual data points are provided in spreadsheet format (xlsx) as Supporting Information. TURBOMOLE is a purchasable software maintained and distributed by the *TURBOMOLE GmbH*. Demo versions are available at https://www.turbomole.org/. The licensed software was provided to us by the bwHPC Cluster JUSTUS2.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jctc.5c00456.

Additional details and figures of the energy evaluation and visualization (PDF)

Detailed table of the individual data points with corresponding adduct formation energies and differences to the reference (XLSX)

AUTHOR INFORMATION

Corresponding Author

Frank M. Boeckler − *Laboratory for Molecular Design & Pharmaceutical Biophysics, Institute of Pharmaceutical Sciences, Department of Pharmacy and Biochemistry and Interfaculty Institute for Biomedical Informatics (IBMI), Eberhard Karls Universität Tübingen, 72076 Tübingen, Germany;* orcid.org/0000-0001-8738-6716; Email: frank.boeckler@uni-tuebingen.de

Authors

Marc U. Engelhardt − *Laboratory for Molecular Design & Pharmaceutical Biophysics, Institute of Pharmaceutical Sciences, Department of Pharmacy and Biochemistry, Eberhard Karls Universität Tübingen, 72076 Tübingen, Germany;* orcid.org/0009-0007-9152-8538

Markus O. Zimmermann − *Laboratory for Molecular Design & Pharmaceutical Biophysics, Institute of Pharmaceutical Sciences, Department of Pharmacy and Biochemistry and Interfaculty Institute for Biomedical Informatics (IBMI), Eberhard Karls Universität Tübingen, 72076 Tübingen, Germany;* orcid.org/0000-0001-6115-8248

Finn Mier − *Laboratory for Molecular Design & Pharmaceutical Biophysics, Institute of Pharmaceutical Sciences, Department of Pharmacy and Biochemistry, Eberhard Karls Universität Tübingen, 72076 Tübingen, Germany*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jctc.5c00456

Author Contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript. F.M.B. and envisioned the research. M.U.E. performed all QM calculations, gathered all results and prepared the corresponding visualizations. M.O.Z. and F.M. contributed to developing the computational strategy and provided comments on the manuscript. M.U.E. prepared the original draft. M.U.E., M.O.Z., F.M., and F.M.B. reviewed, edited, and finalized the manuscript.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors acknowledge support from the state of Baden−Württemberg through bwHPC and the German Research Foundation (DFG) through Grant No. INST 40/575-1 FUGG (JUSTUS 2 cluster).

■ ABBREVIATIONS

XB, halogen bond; X, halogen; QM, quantum mechanical; TPSS, Tao, Perdew, Staroverov, and Scuseria exchange functional; D3, dispersion correction type 3; MP2, Møller−Plesset perturbation method order 2; TZVPP, valence triple-ζ with two sets of polarization functions; TZVPPD, valence triple-ζ with two sets of polarization functions and diffuse function; BSSE, basis set superposition error; cc-pVTZ-PP, correlation-consistent polarized valence triple-ζ function with pseudopotentials; cc-pVQZ-PP, correlation consistent polarized valence quadruple-ζ function with pseudopotentials; B3LYP, Becke, 3-parameter, Lee−Yang−Parr functional; M06−2X, Minnesota 06 hybrid functional with double Hartree−Fock exchange; CCSD(T), coupled cluster with single, double, and perturbative triple excitations; CBS, complete basis set; SCF, self-consistent field; RI, resolution of identity; DFT, density functional theory; ECP, effective core potential; RMSD, root-mean-square deviation; GGA, Generalized Gradient Approximation; SCS-MP2, spin-component-scaled MP2

■ REFERENCES

(1) Bissantz, C.; Kuhn, B.; Stahl, M. A Medicinal Chemist's Guide to Molecular Interactions. *J. Med. Chem.* **2010**, *53* (14), 5061−5084.

(2) Müller-Dethlefs, K.; Hobza, P. Noncovalent Interactions: A Challenge for Experiment and Theory. *Chem. Rev.* **2000**, *100* (1), 143−168.

(3) Anighoro, A. Underappreciated Chemical Interactions in Protein−Ligand Complexes. In *Quantum Mechanics in Drug Discovery*, Heifetz, A., Ed.; Springer: US, 2020; pp 75−86.

(4) Jena, S.; Dutta, J.; Tulsiyan, K. D.; Sahu, A. K.; Choudhury, S. S.; Biswal, H. S. Noncovalent interactions in proteins and nucleic acids: beyond hydrogen bonding and π-stacking. *Chem. Soc. Rev.* **2022**, *51* (11), 4261−4286.

(5) Adhav, V. A.; Saikrishnan, K. The Realm of Unconventional Noncovalent Interactions in Proteins: Their Significance in Structure and Function. *ACS Omega* **2023**, *8* (25), 22268−22284.

(6) Clark, T.; Hennemann, M.; Murray, J. S.; Politzer, P. Halogen bonding: the σ-hole. *J. Mol. Model.* **2007**, *13* (2), 291−296.

(7) Cavallo, G.; Metrangolo, P.; Milani, R.; Pilati, T.; Priimagi, A.; Resnati, G.; Terraneo, G. The Halogen Bond. *Chem. Rev.* **2016**, *116* (4), 2478−2601.

(8) Politzer, P.; Murray, J. S.; Clark, T. Halogen bonding and other σ-hole interactions: a perspective. *Phys. Chem. Chem. Phys.* **2013**, *15* (27), 11178−11189.

(9) Metrangolo, P. R.*Giuseppe Halogen Bonding - Fundamentals and Applications*; Springer, 2008.

(10) Erdélyi, M. Halogen bonding in solution. *Chem. Soc. Rev.* **2012**, *41* (9), 3547−3557.

(11) Politzer, P.; Murray, J. S.; Clark, T. Halogen bonding: an electrostatically-driven highly directional noncovalent interaction. *Phys. Chem. Chem. Phys.* **2010**, *12* (28), 7748−7757.

(12) Desiraju, G. R.; Ho, P. S.; Kloo, L.; Legon, A. C.; Marquardt, R.; Metrangolo, P.; Politzer, P.; Resnati, G.; Rissanen, K. Definition of the halogen bond (IUPAC Recommendations 2013). *Pure Appl. Chem.* **2013**, *85* (8), 1711−1713.

(13) Auffinger, P.; Hays, F. A.; Westhof, E.; Ho, P. S. Halogen bonds in biological molecules. *Proc. Natl. Acad. Sci. U. S. A.* **2004**, *101* (48), 16789−16794.

(14) Andrea, R. V.; Shing Ho, P. The Role of Halogen Bonding in Inhibitor Recognition and Binding by Protein Kinases. *Curr. Top. Med. Chem.* **2007**, *7* (14), 1336−1348.

(15) Parisini, E.; Metrangolo, P.; Pilati, T.; Resnati, G.; Terraneo, G. Halogen bonding in halocarbon−protein complexes: a structural survey. *Chem. Soc. Rev.* **2011**, *40* (5), 2267−2278.

(16) Hardegger, L. A.; Kuhn, B.; Spinnler, B.; Anselm, L.; Ecabert, R.; Stihle, M.; Gsell, B.; Thoma, R.; Diez, J.; Benz, J.; et al. Systematic Investigation of Halogen Bonding in Protein−Ligand Interactions. *Angew. Chem., Int. Ed.* **2011**, *50* (1), 314−318.

(17) Voth, A. R.; Khuu, P.; Oishi, K.; Ho, P. S. Halogen bonds as orthogonal molecular interactions to hydrogen bonds. *Nat. Chem.* **2009**, *1* (1), 74−79.

(18) Parker, A. J.; Stewart, J.; Donald, K. J.; Parish, C. A. Halogen Bonding in DNA Base Pairs. *J. Am. Chem. Soc.* **2012**, *134* (11), 5165−5172.

(19) Xu, Z.; Yang, Z.; Liu, Y.; Lu, Y.; Chen, K.; Zhu, W. Halogen Bond: Its Role beyond Drug−Target Binding Affinity for Drug Discovery and Development. *J. Chem. Inf. Model.* **2014**, *54* (1), 69−78.

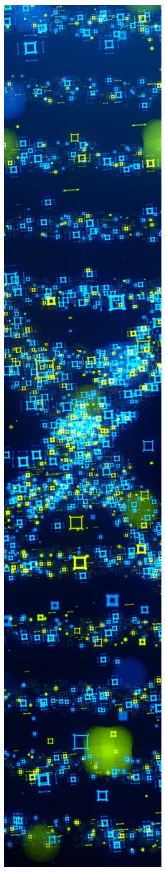
(20) Jiang, L.; Zhang, X.; Zhou, Y.; Chen, Y.; Luo, Z.; Li, J.; Yuan, C.; Huang, M. Halogen bonding for the design of inhibitors by targeting the S1 pocket of serine proteases. *RSC Adv.* **2018**, *8* (49), 28189−28197.

(21) Berger, G.; Frangville, P.; Meyer, F. Halogen bonding for molecular recognition: new developments in materials and biological sciences. *Chem. Commun.* **2020**, *56* (37), 4970−4981.

(22) Dammann, M.; Stahlecker, J.; Zimmermann, M. O.; Klett, T.; Rotzinger, K.; Kramer, M.; Coles, M.; Stehle, T.; Boeckler, F. M. Screening of a Halogen-Enriched Fragment Library Leads to

Unconventional Binding Modes. *J. Med. Chem.* **2022**, *65* (21), 14539−14552.
(23) Vaas, S.; Zimmermann, M. O.; Schollmeyer, D.; Stahlecker, J.; Engelhardt, M. U.; Rheinganz, J.; Drotleff, B.; Olfert, M.; Lämmerhofer, M.; Kramer, M.; et al. Principles and Applications of CF2X Moieties as Unconventional Halogen Bond Donors in Medicinal Chemistry, Chemical Biology, and Drug Discovery. *J. Med. Chem.* **2023**, *66* (15), 10202−10225.
(24) Wilcken, R.; Liu, X.; Zimmermann, M. O.; Rutherford, T. J.; Fersht, A. R.; Joerger, A. C.; Boeckler, F. M. Halogen-Enriched Fragment Libraries as Leads for Drug Rescue of Mutant p53. *J. Am. Chem. Soc.* **2012**, *134* (15), 6810−6818.
(25) Scholfield, M. R.; Zanden, C. M. V.; Carter, M.; Ho, P. S. Halogen bonding (X-bonding): A biological perspective. *Protein Sci.* **2013**, *22* (2), 139−152.
(26) Sirimulla, S.; Bailey, J. B.; Vegesna, R.; Narayan, M. Halogen Interactions in Protein−Ligand Complexes: Implications of Halogen Bonding for Rational Drug Design. *J. Chem. Inf. Model.* **2013**, *53* (11), 2781−2791.
(27) Wilcken, R.; Zimmermann, M. O.; Lange, A.; Joerger, A. C.; Boeckler, F. M. Principles and Applications of Halogen Bonding in Medicinal Chemistry and Chemical Biology. *J. Med. Chem.* **2013**, *56* (4), 1363−1388.
(28) Wilcken, R.; Zimmermann, M. O.; Lange, A.; Zahn, S.; Boeckler, F. M. Using halogen bonds to address the protein backbone: a systematic evaluation. *J. Comput. Aided Mol. Des.* **2012**, *26* (8), 935−945.
(29) Zimmermann, M. O.; Boeckler, F. M. Targeting the protein backbone with aryl halides: systematic comparison of halogen bonding and $\pi\cdots\pi$ interactions using N-methylacetamide. *MedChemComm* **2016**, *7* (3), 500−505.
(30) Wilcken, R.; Zimmermann, M. O.; Lange, A.; Zahn, S.; Kirchner, B.; Boeckler, F. M. Addressing Methionine in Molecular Design through Directed Sulfur−Halogen Bonds. *J. Chem. Theory Comput.* **2011**, *7* (7), 2307−2315.
(31) Lange, A.; Zimmermann, M. O.; Wilcken, R.; Zahn, S.; Boeckler, F. M. Targeting Histidine Side Chains in Molecular Design through Nitrogen−Halogen Bonds. *J. Chem. Inf. Model.* **2013**, *53* (12), 3178−3189.
(32) Zimmermann, M. O.; Lange, A.; Zahn, S.; Exner, T. E.; Boeckler, F. M. Using Surface Scans for the Evaluation of Halogen Bonds toward the Side Chains of Aspartate, Asparagine, Glutamate, and Glutamine. *J. Chem. Inf. Model.* **2016**, *56* (7), 1373−1383.
(33) Engelhardt, M. U.; Zimmermann, M. O.; Dammann, M.; Stahlecker, J.; Poso, A.; Kronenberger, T.; Kunick, C.; Stehle, T.; Boeckler, F. M. Halogen Bonding on Water—A Drop in the Ocean? *J. Chem. Theory Comput.* **2024**, *20* (23), 10716−10730.
(34) Zhou, P.; Lv, J.; Zou, J.; Tian, F.; Shang, Z. Halogen−water−hydrogen bridges in biomolecules. *J. Struct. Biol.* **2010**, *169* (2), 172−182.
(35) Matter, H.; Nazaré, M.; Güssregen, S.; Will, D. W.; Schreuder, H.; Bauer, A.; Urmann, M.; Ritter, K.; Wagner, M.; Wehner, V. Evidence for C-Cl/C-Br$\cdots\pi$ Interactions as an Important Contribution to Protein−Ligand Binding Affinity. *Angew. Chem., Int. Ed.* **2009**, *48* (16), 2911−2916.
(36) Shah, M. B.; Liu, J.; Zhang, Q.; Stout, C. D.; Halpert, J. R. Halogen−π Interactions in the Cytochrome P450 Active Site: Structural Insights into Human CYP2B6 Substrate Selectivity. *ACS Chem. Biol.* **2017**, *12* (5), 1204−1210.
(37) Heroven, C.; Georgi, V.; Ganotra, G. K.; Brennan, P.; Wolfreys, F.; Wade, R. C.; Fernández-Montalván, A. E.; Chaikuad, A.; Knapp, S. Halogen−Aromatic π Interactions Modulate Inhibitor Residence Times. *Angew. Chem., Int. Ed.* **2018**, *57* (24), 7220−7224.
(38) Wallnoefer, H. G.; Fox, T.; Liedl, K. R.; Tautermann, C. S. Dispersion dominated halogen−π interactions: energies and locations of minima. *Phys. Chem. Chem. Phys.* **2010**, *12* (45), 14941−14949.
(39) Forni, A.; Pieraccini, S.; Rendine, S.; Gabas, F.; Sironi, M. Halogen-Bonding Interactions with π Systems: CCSD(T), MP2, and DFT Calculations. *ChemPhysChem* **2012**, *13* (18), 4224−4234.
(40) Wolters, L. P.; Schyman, P.; Pavan, M. J.; Jorgensen, W. L.; Bickelhaupt, F. M.; Kozuch, S. The many faces of halogen bonding: a review of theoretical models and methods. *WIREs Comput. Mol. Sci.* **2014**, *4* (6), 523−540.
(41) Kozuch, S.; Martin, J. M. L. Halogen Bonds: Benchmarks and Theoretical Analysis. *J. Chem. Theory Comput.* **2013**, *9* (4), 1918−1931.
(42) Ford, M. C.; Ho, P. S. Computational Tools To Model Halogen Bonds in Medicinal Chemistry. *J. Med. Chem.* **2016**, *59* (5), 1655−1670.
(43) Youn, I. S.; Kim, D. Y.; Cho, W. J.; Madridejos, J. M. L.; Lee, H. M.; Kołaski, M.; Lee, J.; Baig, C.; Shin, S. K.; Filatov, M.; et al. Halogen−π Interactions between Benzene and X2/CX4 (X = Cl, Br): Assessment of Various Density Functionals with Respect to CCSD(T). *J. Phys. Chem. A* **2016**, *120* (46), 9305−9314.
(44) Řezáč, J.; Riley, K. E.; Hobza, P. Benchmark Calculations of Noncovalent Interactions of Halogenated Molecules. *J. Chem. Theory Comput.* **2012**, *8* (11), 4285−4292.
(45) Zhu, Z.; Xu, Z.; Zhu, W. Interaction Nature and Computational Methods for Halogen Bonding: A Perspective. *J. Chem. Inf. Model.* **2020**, *60* (6), 2683−2696.
(46) Young, D. C. *Computational Chemistry: A Practical Guide for Applying Techniques to Real World Problems*; Wiley, 2001.
(47) Jensen, F. *Introduction to Computational Chemistry*; Wiley, 2017.
(48) Møller, C.; Plesset, M. S. Note on an Approximation Treatment for Many-Electron Systems. *Phys. Rev.* **1934**, *46* (7), 618−622.
(49) Head-Gordon, M.; Pople, J. A.; Frisch, M. J. MP2 energy evaluation by direct methods. *Chem. Phys. Lett.* **1988**, *153* (6), 503−506.
(50) Pople, J. A.; Binkley, J. S.; Seeger, R. Theoretical models incorporating electron correlation. *Int. J. Quantum Chem.* **1976**, *10* (S10), 1−19.
(51) Pople, J. A.; Seeger, R.; Krishnan, R. Variational configuration interaction methods and comparison with perturbation theory. *Int. J. Quantum Chem.* **1977**, *12* (S11), 149−163.
(52) Riley, K. E.; Řezáč, J.; Hobza, P. MP2.X: a generalized MP2.5 method that produces improved binding energies with smaller basis sets. *Phys. Chem. Chem. Phys.* **2011**, *13* (47), 21121−21125.
(53) Grimme, S. Improved second-order Møller−Plesset perturbation theory by separate scaling of parallel- and antiparallel-spin pair correlation energies. *J. Chem. Phys.* **2003**, *118* (20), 9095−9102.
(54) Grimme, S.; Goerigk, L.; Fink, R. F. Spin-component-scaled electron correlation methods. *WIREs Comput. Mol. Sci.* **2012**, *2* (6), 886−906.
(55) Kohn, W.; Sham, L. J. Self-Consistent Equations Including Exchange and Correlation Effects. *Phys. Rev.* **1965**, *140* (4A), A1133−A1138.
(56) Staroverov, V. N.; Scuseria, G. E.; Tao, J.; Perdew, J. P. Comparative assessment of a new nonempirical density functional: Molecules and hydrogen-bonded complexes. *J. Chem. Phys.* **2003**, *119* (23), 12129−12137.
(57) Perdew, J. P.; Chevary, J. A.; Vosko, S. H.; Jackson, K. A.; Pederson, M. R.; Singh, D. J.; Fiolhais, C. Atoms, molecules, solids, and surfaces: Applications of the generalized gradient approximation for exchange and correlation. *Phys. Rev. B* **1992**, *46* (11), 6671−6687.
(58) Becke, A. D. Density-functional thermochemistry. III. The role of exact exchange. *J. Chem. Phys.* **1993**, *98* (7), 5648−5652.
(59) Lee, C.; Yang, W.; Parr, R. G. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Phys. Rev. B* **1988**, *37* (2), 785−789.
(60) Zhao, Y.; Truhlar, D. G. The M06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: two new functionals and systematic testing of four M06-class functionals and 12 other functionals. *Theor. Chem. Acc.* **2008**, *120* (1), 215−241.
(61) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *J. Chem. Phys.* **2010**, *132* (15), 154104.

(62) Schäfer, A.; Huber, C.; Ahlrichs, R. Fully optimized contracted Gaussian basis sets of triple zeta valence quality for atoms Li to Kr. *J. Chem. Phys.* **1994**, *100* (8), 5829−5835.
(63) Dunning, T. H., Jr. Gaussian basis sets for use in correlated molecular calculations. I. The atoms boron through neon and hydrogen. *J. Chem. Phys.* **1989**, *90* (2), 1007−1023.
(64) Woon, D. E.; Dunning, T. H., Jr. Gaussian basis sets for use in correlated molecular calculations. IV. Calculation of static electrical response properties. *J. Chem. Phys.* **1994**, *100* (4), 2975−2988.
(65) Halkier, A.; Helgaker, T.; Jørgensen, P.; Klopper, W.; Koch, H.; Olsen, J.; Wilson, A. K. Basis-set convergence in correlated calculations on Ne, N2, and H2O. *Chem. Phys. Lett.* **1998**, *286* (3), 243−252.
(66) Boys, S. F.; Bernardi, F. The calculation of small molecular interactions by the differences of separate total energies. Some procedures with reduced errors. *Mol. Phys.* **1970**, *19* (4), 553−566.
(67) Alvarez-Idaboy, J. R.; Galano, A. Counterpoise corrected interaction energies are not systematically better than uncorrected ones: comparison with CCSD(T) CBS extrapolated values. *Theor. Chem. Acc.* **2010**, *126* (1), 75−85.
(68) Pitoňák, M.; Neogrády, P.; Černý, J.; Grimme, S.; Hobza, P. Scaled MP3 Non-Covalent Interaction Energies Agree Closely with Accurate CCSD(T) Benchmark Data. *ChemPhysChem* **2009**, *10* (1), 282−289.
(69) Riley, K. E.; Řezáč, J.; Hobza, P. The performance of MP2.5 and MP2.X methods for nonequilibrium geometries of molecular complexes. *Phys. Chem. Chem. Phys.* **2012**, *14* (38), 13187−13193.
(70) Walker, M.; Harvey, A. J. A.; Sen, A.; Dessent, C. E. H. Performance of M06, M06−2X, and M06-HF Density Functionals for Conformationally Flexible Anionic Clusters: M06 Functionals Perform Better than B3LYP for a Model System with Dispersion and Ionic Hydrogen-Bonding Interactions. *J. Phys. Chem. A* **2013**, *117* (47), 12590−12600.
(71) Sethio, D.; Raggi, G.; Lindh, R.; Erdélyi, M. Halogen Bond of Halonium Ions: Benchmarking DFT Methods for the Description of NMR Chemical Shifts. *J. Chem. Theory Comput.* **2020**, *16* (12), 7690−7701.
(72) TURBOMOLE V7.7.1 2019, a development of University of Karlsruhe and Forschungszentrum Karlsruhe GmbH, 1989−2007, *TURBOMOLE GmbH*, since **2007**.
(73) Weigend, F.; Häser, M. RI-MP2: first derivatives and global consistency. *Theor. Chem. Acc.* **1997**, *97* (1), 331−340.
(74) Weigend, F.; Häser, M.; Patzelt, H.; Ahlrichs, R. RI-MP2: optimized auxiliary basis sets and demonstration of efficiency. *Chem. Phys. Lett.* **1998**, *294* (1), 143−152.
(75) Hättig, C.; Weigend, F. CC2 excitation energy calculations on large molecules using the resolution of the identity approximation. *J. Chem. Phys.* **2000**, *113* (13), 5154−5161.
(76) Hoffmann, R. An Extended Hückel Theory. I. Hydrocarbons. *J. Chem. Phys.* **2004**, *39* (6), 1397−1412.
(77) Weigend, F.; Ahlrichs, R. Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy. *Phys. Chem. Chem. Phys.* **2005**, *7* (18), 3297−3305.
(78) Häser, M.; Ahlrichs, R. Improvements on the direct SCF method. *J. Comput. Chem.* **1989**, *10* (1), 104−111.
(79) Feyereisen, M.; Fitzgerald, G.; Komornicki, A. Use of approximate integrals in ab initio theory. An application in MP2 energy calculations. *Chem. Phys. Lett.* **1993**, *208* (5), 359−363.
(80) Hättig, C. Geometry optimizations with the coupled-cluster model CC2 using the resolution-of-the-identity approximation. *J. Chem. Phys.* **2003**, *118* (17), 7751−7761.
(81) Hättig, C.; Hellweg, A.; Köhn, A. Distributed memory parallel implementation of energies and gradients for second-order Møller−Plesset perturbation theory with the resolution-of-the-identity approximation. *Phys. Chem. Chem. Phys.* **2006**, *8* (10), 1159−1169.
(82) *The PyMOL Molecular Graphics System, Version 2.5.5*; Schrödinger, LLC, 2015.
(83) *bwForCluster − JUSTUS 2*. https://wiki.bwhpc.de/e/JUSTUS2 (accessed 2024−2025).



https://doi.org/10.1021/acs.jctc.5c00456
J. Chem. Theory Comput. 2025, 21, 6174−6183

Appendix C: Publication 3

A QM-AI Approach for the Acceleration of Accurate Assessments of Halogen-π Interactions by Training Neural Networks

Marc U. Engelhardt, Finn Mier, Markus O. Zimmermann, Frank M. Boeckler;

Journal of Chemical Information and Modeling, 2025

DOI: 10.1021/acs.jcim.5c02136

JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

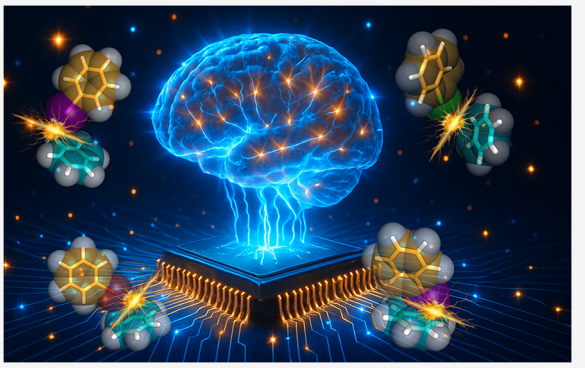
pubs.acs.org/jcim
Article

A QM-AI Approach for the Acceleration of Accurate Assessments of Halogen-π Interactions by Training Neural Networks

Marc U. Engelhardt, Finn Mier, Markus O. Zimmermann, and Frank M. Boeckler*

Cite This: https://doi.org/10.1021/acs.jcim.5c02136

ABSTRACT: Noncovalent interactions, such as halogen bonds (XB), play a crucial role in molecular recognition and drug design, yet halogen···π contacts remain comparatively underexplored. Here, we report a proof-of-concept QM-AI approach that integrates high-level quantum mechanical (QM) calculations with neural networks (NNs) to predict halogen···π interaction energies. Nearly 1.4 million MP2/TZVPP single-point calculations on halobenzene−benzene complexes were carried out to generate exhaustive training data, which were represented by simple geometric descriptors as input features for machine learning. The resulting neural network model is specifically designed to capture σ-hole-driven halogen···π interactions under well-defined geometric constraints. The resulting model reproduced reference interaction energies with excellent accuracy (R^2 = 0.998, RMSE = 0.16 kJ/mol) and maintained strong performance on independent, randomly generated and PDB-derived test sets. Previously, we have demonstrated in a benchmarking study that "gold standard" CCSD(T) energies of this interaction can be appropriately represented by MP2/TZVPP calculations, but at a better calculation efficiency by 2 orders of magnitude ($\sim 10^2$). Consequently, we herein exploit a methodological "extension" from CCSD(T) → MP2 → NNs. Our approach maintains accuracy close to CCSD(T) benchmarks while achieving a runtime acceleration of up to 8 orders of magnitude ($\sim 10^8$) compared to MP2 calculations. This study demonstrates the feasibility of fast, accurate neural network models based on QM data for halogen···π interactions in a QM-AI approach.



INTRODUCTION

Noncovalent interactions are central to a wide range of processes in biological systems and molecular recognition, such as protein folding, drug binding, and protein−ligand interactions. In particular, understanding the individual contributions is crucial for elucidating biomolecular functions and guiding the rational design of potential therapeutics.[1−5] Among diverse types of noncovalent interactions, halogen bonding (XB) has emerged as a powerful and versatile tool due to its high directionality and ability to interact with a multitude of different partners in the protein binding site. It is defined by a directional attraction between a nucleophilic site and an electrophilic region, the σ-hole, typically located on the elongation of the R−X axis on a halogen atom such as chlorine, bromine, or iodine.[6−11] The highly anisotropic electron distribution around halogen atoms results in a significant lateral electron density, oriented perpendicular to the R−X bond axis. This feature contributes to the characteristically pronounced directionality of halogen bond interactions. In systems where the substituent R exerts a strong electron-withdrawing effect on the halogen (X), such "tuning effects" can significantly enhance the strength of the halogen bond.[12−15] Accordingly, the strategic incorporation of halogen bonds in drug design has attracted increasing attention to optimize pharmacological profiles by enhancing ligand binding affinity or improving specificity and stability in protein−ligand complexes.[16−33]

Computational methods have been extensively applied to accurately characterize the geometric and energetic features of halogen bonding. To date, halogen bonds have been systematically investigated in the context of various nucleophilic moieties within binding sites, including backbone carbonyl groups and the π-surface of the peptide bond,[34,35] the sulfur atom in methionine,[36] the nitrogen atoms in histidine,[37] carboxylate groups of aspartate and glutamate, as well as carboxamide groups of asparagine and glutamine moieties,[38] and the oxygen atoms of water molecules.[39,40] In contrast, halogen bonding involving the electron-rich π-system of the aromatic side chain residues of phenylalanine, tyrosine, histidine, and tryptophan, however, remains comparatively unexplored in the context of protein−ligand interactions.[41−45] In a very recent study, we used quantum mechanical

Received: September 3, 2025
Revised: November 11, 2025
Accepted: November 13, 2025

https://doi.org/10.1021/acs.jcim.5c02136
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

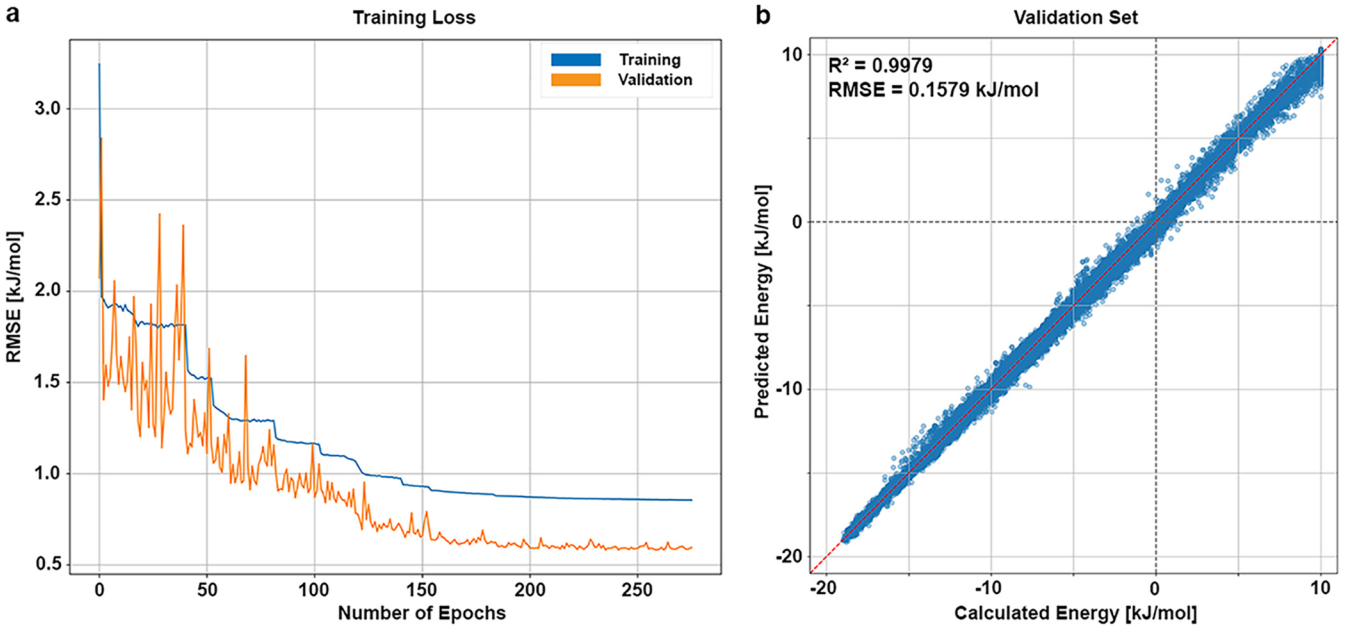


Figure 1. Results of the training process of the final model. (a) RMSE values in kJ/mol for each training epoch. The blue values represent the RMSE for each epoch of the training process, while orange values represent the RMSE of the validation after each training epoch. (b) Final model performance on the validation set with a coefficient of determination of R_{val}^2 = 0.9979, and root-mean-square error of $RMSE_{val}$ = 0.1579 kJ/mol. The calculated energy is plotted against the predicted energy. The red, dashed line indicates the perfect correlation between calculated and predicted value, while the gray dashed lines indicate the transition from negative to positive energies.

calculations to examine halogen···π interactions between halobenzenes (strongly focusing on iodine interactions) as ligand systems and benzene as a surrogate for the aromatic side chain of phenylalanine.[46] We compared calculations of different QM methods and basis sets and showed that calculations on an MP2 level of theory using a reasonably large basis set of TZVPP or cc-pVTZ yield highly accurate energies compared to reference calculations (CCSD(T)/$CBS_{extrapolation}$) while maintaining feasible computational demands, even for larger data sets. This transition from CCSD(T) to MP2 corresponds to a computational speed-up of about 2 orders of magnitude ($\sim 10^2$) without a significant loss in accuracy. This enabled us to generate the basis for the following machine learning approach, aiming for a tremendous speed-up to an almost instant geometric assessment of high quality energies of halogen···π interactions.

The computational complexity of QM methods can become a significant obstacle, when applied to larger systems or when evaluating large data sets. To address this limitation, data-driven approaches such as neural networks (NNs) are emerging as powerful tools to complement or even replace QM calculations in certain contexts.[47−50] One of the key strengths of NN-based approaches is their scalability and versatility. Once trained, models can predict interaction energies for a wide range of molecular systems in a tiny fraction of the time required for QM calculations. Inputs to these NNs are typically molecular descriptors representing crucial geometric features such as interaction distances and angles. Thus, the integration of QM-calculated data sets with NN-based models represents a promising approach to bridge the gap between computational cost and accuracy.[51−53]

Shaw et al.[54] introduced a simple two-parameter statistical model to predict halogen bond interaction energies for a small data set of halogenated compounds. Using this basic type of machine learning model on an unseen test set of 80 complexes, they achieve accuracies within ~2.1 kJ/mol emphasizing the applicability of machine learning in general. In 2023, Samuel et al.[55] published a general overview of various machine learning approaches on halogen bonding and their differences concluding that it is a "...powerful tool for unravelling the intricacies of molecular interactions and guiding the design of functional molecular systems". They foresee continued progress as larger data sets become more available, and hybrid quantum-machine learning ("QM-AI") approaches become more prevalent. Devore et al.[56] published a machine learning-based approach to characterize halogen bonding interactions using molecular fingerprints as input descriptors. Their study showed that supervised learning models can accurately classify halogen-bond donor types and predict interaction properties, further highlighting the potential of data-driven techniques to complement QM methods in evaluating halogen bonding complexes.

In this study, we focus on the systematic investigation of halogen···π interactions using high-level quantum mechanical (QM) methods and integrating them into NN models. Based on findings of our previous study, a data set of interaction geometries composed of halobenzenes (chlorobenzene, bromobenzene, and iodobenzene) in complex with benzene is generated and used in single point calculations on an MP2/TZVPP level of theory. To specifically investigate and assess σ-hole interactions, a systematic grid of interaction geometries was generated in planes parallel to the π-plane. Each plane was positioned at a fixed halogen···π-plane separation, with distances ranging from $d_{min(X\cdots\pi\text{-plane})}$ = 2.75 Å to $d_{max(X\cdots\pi\text{-plane})}$ = 4.50 Å. In this distance definition ($d_{X\cdots\pi\text{-plane}}$) the respective point on the π-plane is individually determined by the normal onto the π-plane through the halogen atom. Within each plane, geometries were sampled to include orientations deviating by no more than 40° between the C−X bond vector and the normal onto the benzene molecule (π-plane). This setup

https://doi.org/10.1021/acs.jcim.5c02136
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

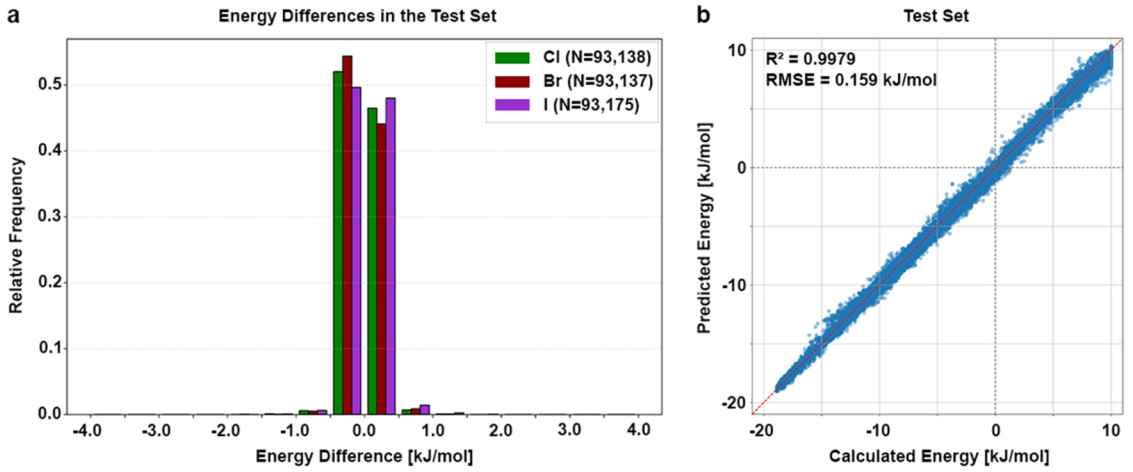


Figure 2. Model performance on the test set. (a) The histogram shows the relative frequencies of energy differences between calculated and predicted energy (total of 279,450 data points) in bins of 0.5 kJ/mol from −4.0 to 4.0 kJ/mol for chlorine (green, N = 93,138), bromine (dark red, N = 93,137), and iodine (purple, N = 93,175) separately. Energy differences are calculated as $\Delta\Delta E = \Delta E_{calc} - \Delta E_{pred}$. Larger values are clipped to the respective limitation for better visibility. (b) Calculated energy is plotted against the predicted energy. The model achieved an R^2 = 0.9979 with an RMSE = 0.159 kJ/mol. The red, dashed line indicates the perfect correlation between calculated and predicted value, while the gray dashed lines indicate the transition from negative to positive energies.

ensures that the data set selectively represents configurations characteristic of σ-hole interactions, while reducing contributions of secondary interactions, e.g $\pi\cdots\pi$, or C−H$\cdots\pi$ interactions. The results are then used to train NN models in a supervised learning approach to predict adduct formation energies with high accuracy and significantly reduced computational cost. We demonstrated that MP2 reproduces CCSD(T) reference data with excellent accuracy. This close agreement indicates that MP2 captures the essential properties of these noncovalent interactions, while offering a substantial computational speed-up of approximately 2 orders of magnitude (10^2) compared to CCSD(T). Building on this finding, we propose a methodological "double jump" approach, first from CCSD(T) to MP2, and then from MP2 to machine-learning (NN) models. In the first "jump", CCSD(T) → MP2, we replace the computationally prohibitive CCSD(T) calculations with MP2, achieving near-CCSD(T) accuracy at a fraction of the cost. In the second "jump", MP2 → NNs, we train neural network models on extensive MP2 data sets, enabling the prediction of interaction energies with a further acceleration of approximately 8 orders of magnitude (10^8) compared to the underlying MP2 computations. Together, this "double jump" preserves the accuracy hierarchy established between CCSD(T) and MP2, while extending it further into the field of machine learning. As a result, the NN models inherit the near-CCSD(T) accuracy of MP2 but with unprecedented computational efficiency. A PDB[57] scan provided examples from real-world biological systems to evaluate the model and its potential to quickly predict halogen bond interaction energies. By integrating QM calculations with machine learning techniques, this work serves as an initial step to set the stage for the generation of subsequent QM-AI hybrid models able to assess interactions between halogens and other acceptor systems. Ultimately, we aim to employ these models into the molecular docking framework PLANTS, to enhance the identification and scoring of halogen bond interactions in protein−ligand complexes.

RESULTS AND DISCUSSION

NN Model Training and Validation. First, we conducted almost 1.4 million single-point calculations of the generated interaction geometries. Energies were calculated using the supermolecular approach. Features were derived from the individual interaction geometries. A complete list of all features with their descriptions, along with a more detailed schematic of the feature definitions, is provided in the Supporting Information (Figure S1 and Table S1). To ensure the feature space is translationally and rotationally independent from the underlying coordinate system, we chose pairwise distances and angles. The resulting data set was partitioned into training, validation, and test subsets (Figure 8). Final model training was conducted using the training subset, with performance evaluation on the validation subset at the end of each epoch. Figure 1a illustrates the progression of model performance on the training and validation data set over consecutive epochs. Training was terminated after 276 epochs upon satisfaction of the early stopping criterion. The results of the final evaluation step on the validation set are presented in Figure 1b, where the predicted adduct formation energies are plotted against the calculated values. The model achieved a coefficient of determination R^2 = 0.9979 and a root-mean-square error $RMSE_{val}$ = 0.1579 kJ/mol, indicating excellent predictive accuracy. Results on the test set also show excellent accuracies with minimal amounts of large energy differences $\Delta\Delta E$ (Figure 2a). Results are given for each halogen individually. Most differences lie within ±0.5 kJ/mol with only a minimal fraction of 1.8% in total lying beyond. An overall maximum of $\Delta\Delta E$ = +2.40 kJ/mol and a minimum of $\Delta\Delta E$ = −2.35 kJ/mol was observed. Overall, the model achieved an R^2 = 0.9979 and an $RMSE_{test}$ = 0.1590 kJ/mol on the test set, indicating that the predictive accuracy is maintained even on previously unseen data (Figure 2b). The observed performance may be inflated due to similarities between the training and test sets, potentially introducing data leakage as the data points likely reside within the same feature space. Consequently, a more profound evaluation involves assessing the model's general-

https://doi.org/10.1021/acs.jcim.5c02136
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

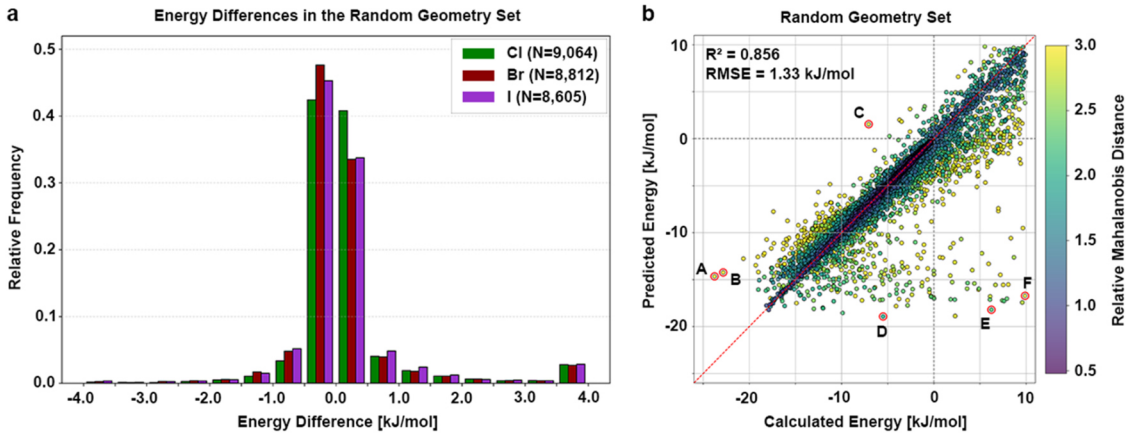


Figure 3. Model performance on the random geometry set. (a) The histogram shows the relative frequencies of energy differences between calculated and predicted energy (total of 26,481 data points) in bins of 0.5 kJ/mol from −4.0 to 4.0 kJ/mol for chlorine (green, N = 9064), bromine (dark red, N = 8812), and iodine (purple, N = 8605) separately. Energy differences are calculated as $\Delta\Delta E = \Delta E_{calc} - \Delta E_{pred}$. Larger values are clipped to the respective limitation for better visibility. (b) Calculated energy is plotted against the predicted energy. The model achieved an R^2 = 0.856 with an RMSE = 1.33 kJ/mol. The red, dashed line indicates the perfect correlation between calculated and predicted value, while the gray dashed lines indicate the transition from negative to positive energies. Each data point is colored according to its relative MD with respect to the given color scale. Data points outlined with a red circle and labeled with (A−F) are shown in detail in Figure 4.

ization capability on entirely unknown data that fall outside the distribution of the training set.

The geometric inclusion criteria employed in the subsequently used data sets of randomly generated and PDB-derived geometries (next sections) were defined with relatively broad angular and distance limits. This approach was chosen to capture a diverse range of halogen···π interaction geometries, including those near the characteristic boundaries of σ-hole interactions. While more restrictive definitions would exclude such borderline geometries and reduce statistical outliers, they would also narrow the physical diversity of the data set and potentially obscure the model's limitations. Our aim is to evaluate the model's generalizability across a wide spectrum of interaction geometries, rather than to optimize its performance under tightly constrained conditions. This broader definition also preserves the size and representativeness of the experimentally derived PDB data set, which would otherwise have been substantially reduced by stricter geometric filters applied to the already small data set.

Prediction of Adduct Formation Energies on Unseen Data. A data set of interaction geometries with randomly chosen translational and rotational features was generated. For a total data set of 30,000 geometries, single point calculations on the MP2/TZVPP level of theory were conducted. Positive energies of more than +10 kJ/mol (repulsions) were excluded, resulting in a final set of 26,481 complexes. Geometric features were extracted from each complex and fed into the model. A detailed table incorporating the individual data points, energies E_{calc} and E_{pred}, as well as the difference between both is provided in spreadsheet format in the Supporting Information. The distribution of energy differences between the calculated and predicted energies is illustrated in Figure 3a. Results are shown for each halogen individually. Most differences still lie within ±0.5 kJ/mol. Approximately 14% of the data points lie beyond ±0.5 kJ/mol, with a maximum of $\Delta\Delta E$ = +27.02 kJ/mol and a minimum $\Delta\Delta E$ = −9.62 kJ/mol. Such outliers are of high interest, because the question arises where they originate from and why the model's prediction on these data points is poor. Different possibilities can be considered when analyzing outliers in a model's predictions: (i) The features of the outlier lie outside the range covered by the training data. In this case, poor predictive performance may indicate that the model lacks generalizability and is unable to extrapolate to previously unseen feature spaces. (ii) The features of the outlier lie within the distribution of the training data, but the model still fails to predict accurately. This suggests a deficiency in the model itself, possibly due to overfitting, underfitting, or an insufficiently expressive model architecture. This raises concerns about the model's internal representation and robustness. To further investigate these outliers and examine their deviation from the expected model behavior, a scatter plot of the calculated and the predicted energy values was generated (Figure 3b). The coefficient of determination R^2 = 0.856 and the root-mean-square error RMSE = 1.33 kJ/mol confirm the model's overall strong accuracy. The plot highlights individual deviations across the test set with a color scale indicating the relative Mahalanobis distance (MD), an estimate of how far a given data point lies from the center of the feature distribution. The MD accounts for correlations between input features and scales the distance based on the covariance structure of the data, making it well-suited to identify outliers in multivariate spaces. In this context, a higher MD suggests that the corresponding data point differs substantially from the typical feature profiles seen during training. The majority of the data points lie close to the diagonal, indicating good agreement between predicted and calculated values and consistent with the narrow energy differences observed in Figure 3a. Notably, data points that deviate more strongly from the parity line also tend to exhibit higher MD, suggesting a reasonable correlation between poor predictive performance and a greater dissimilarity from the training data distribution. Additionally, several data points (labeled A−F) show significant deviation and are marked as prominent outliers for illustrative purposes. These points serve as examples for the error sources discussed above and are looked at in detail. Calculated and predicted adduct formation energies, energy differences, distances $d_{X\cdots\pi\text{-plane}}$ between the halogen and the π-plane, as well as the angle between the C−X vector and the normal to the π-plane $\alpha_{C-X\cdots\perp(\pi\text{-plane})}$ of these examples are listed in Table 1. Figure 4 visualizes the

https://doi.org/10.1021/acs.jcim.5c02136
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Table 1. Overview of Energy Values and Geometric Parameters for Outlier Structures (A–F) (from the Random Geometry Set) Highlighted in Figure 3 and Depicted in Figure 4[a]

	halogen	ΔE_{calc} [kJ/mol]	ΔE_{pred} [kJ/mol]	$\Delta\Delta E$ [kJ/mol]	$d_{X\cdots\pi\text{-plane}}$ [Å]	$\alpha_{C-X\cdots\perp(\pi\text{-plane})}$ [deg]
A	I	−23.77	−14.66	−9.11	1.70	59.18
B	I	−22.83	−14.24	−8.59	1.53	59.66
C	Br	−7.06	1.54	−8.60	1.83	57.28
D	I	−5.50	−18.93	13.43	2.73	59.69
E	I	6.24	−18.21	24.45	2.45	57.69
F	Cl	9.89	−16.73	26.62	1.76	53.82

[a]The halogen symbolizes the interacting halobenzene. Values of calculated and predicted energies, as well as the difference between both are given in kJ/mol. Distance values $d_{X\cdots\pi\text{-plane}}$ between the halogen and the π-plane are given in Å. Angle values between the C–X vector and the normal to the π-plane are given in degree [deg].

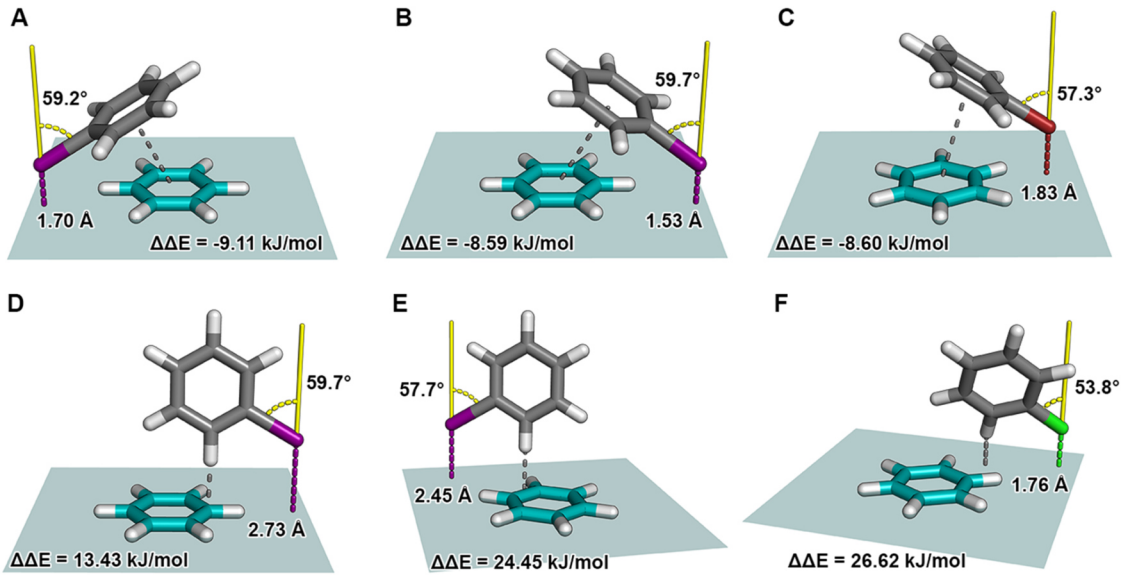


Figure 4. Interaction geometries corresponding to selected outliers of the random geometry set identified in the scatter plot shown in Figure 3. These structures represent data points with high deviations between calculated and predicted adduct formation energies. Each geometry (A–F) illustrates the spatial arrangement of the halobenzene (gray) relative to the π-system of benzene (teal), including the halogen distance $d_{X\cdots\pi\text{-plane}}$ (dashed line, colored according to the halogen) to the π-system plane in Å, the torsion angle $\alpha_{C-X\cdots\perp(\pi\text{-plane})}$ (colored yellow) in degrees between the C–X vector and the normal of the π-system plane, as well as the energy difference $\Delta\Delta E$. The gray dashed line highlights the type of interaction that may contribute to the observed prediction error. For better visibility, the teal-colored plane illustrates the benzene plane and the dimensions of the model's training grid. (A) Iodobenzene interaction with $d_{I\cdots\pi\text{-plane}}$ = 1.70 Å and $\alpha_{C-I\cdots\perp(\pi\text{-plane})}$ = 59.2°. The gray dashed line indicates a $\pi\cdots\pi$-interaction. (B) Iodobenzene interaction with $d_{I\cdots\pi\text{-plane}}$ = 1.53 Å and $\alpha_{C-I\cdots\perp(\pi\text{-plane})}$ = 59.7°. The gray dashed line indicates a $\pi\cdots\pi$-interaction. (C) Bromobenzene interaction with $d_{Br\cdots\pi\text{-plane}}$ = 1.83 Å and $\alpha_{C-Br\cdots\perp(\pi\text{-plane})}$ = 57.3°. The gray dashed line indicates a $\pi\cdots\pi$ interaction. (D) Iodobenzene interaction with $d_{I\cdots\pi\text{-plane}}$ = 2.73 Å and $\alpha_{C-I\cdots\perp(\pi\text{-plane})}$ = 59.7° The gray dashed line indicates a C–H$\cdots\pi$ contact. (E) Iodobenzene interaction with $d_{I\cdots\pi\text{-plane}}$ = 2.45 Å and $\alpha_{C-I\cdots\perp(\pi\text{-plane})}$ = 57.7°. The gray dashed line indicates a C–H$\cdots\pi$ contact. (F) Chlorobenzene interaction with $d_{Cl\cdots\pi\text{-plane}}$ = 1.76 Å and $\alpha_{C-Cl\cdots\perp(\pi\text{-plane})}$ = 53.8° The gray dashed line indicates a C–H$\cdots\pi$ contact.

corresponding interaction geometries. The halogen$\cdots\pi$-interactions in Figure 4A–C all feature geometries, where the halogen lies near the edge of the model's training grid dimensions (illustrated as the teal-colored plane through the benzene), with a C–X vector pointing away from the π-plane, while engaging in potential $\pi\cdots\pi$ interactions. The geometry of example A shows an iodobenzene molecule in complex with benzene and is characterized by a short distance of $d_{I\cdots\pi\text{-plane}}$ = 1.70 Å and an angle of $\alpha_{C-I\cdots\perp(\pi\text{-plane})}$ = 59.2°. A large energy difference $\Delta\Delta E$(A) = −9.11 kJ/mol (ΔE_{calc} = −23.77 kJ/mol, ΔE_{pred} = −14.66 kJ/mol) indicates that the model significantly underestimates the interaction energy. Geometric features likely fall outside the typical range of the feature space covered by the training, as indicated by an MD more than three times higher than that of the feature distribution center. A direct comparison with the training data supports this assumption. The distance in example A $d_{I\cdots\pi\text{-plane}}$(A) = 1.70 Å is substantially shorter than the minimum value in the training set $d_{min(X\cdots\pi\text{-plane})}$(training) = 2.75 Å, and the angle $\alpha_{C-I\cdots\perp(\pi\text{-plane})}$(A) = 59.2° exceeds the maximum training value of $\alpha_{max(C-X\cdots\perp(\pi\text{-plane}))}$(training) = 40°. Furthermore, the halogen's proximity to the benzene plane, and thus its in-plane hydrogen atoms, may also contribute beneficially due to contacts between the negatively charged belt of the halogen atom and nearby hydrogen atoms. Similar observations can be found in B and C with underestimations of the interaction energy for both examples ($\Delta\Delta E$(B) = −8.59 kJ/mol, $\Delta\Delta E$(C) = −8.6 kJ/mol). Distance and angle values also show significant deviations from the training features with similar geometric attributes. Such errors suggest that the model lacks prior information about $\pi\cdots\pi$ interaction motifs or X$\cdots$H motifs with such compact and tilted geometries, leading to a systematic underestimation of their energetically favorable contribution. Looking at examples D–F, the spatial arrange-

https://doi.org/10.1021/acs.jcim.5c02136
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

ments change from $\pi\cdots\pi$ to C−H$\cdots\pi$ contributions. With positive energy differences of $\Delta\Delta E$(D) = 13.43 kJ/mol, $\Delta\Delta E$(E) = 24.45 kJ/mol, and $\Delta\Delta E$(F) = 26.62 kJ/mol, the model significantly overestimates these interactions. Halogen atom positions lie close to the edges of the grid dimensions while the halobenzene is oriented above the benzene. While the distance of example D, $d_{\text{I}\cdots\pi\text{-plane}}$(D) = 2.73 Å, shows direct proximity to the minimum training distance, the angle value $\alpha_{\text{C-I}\cdots\perp(\pi\text{-plane})}$(D)= 59.7° still deviates significantly from the training set. However, the acceptable distance, combined with a reasonable C−H$\cdots\pi$ contact yields a calculated energy of ΔE_{calc}(D)= −5.5 kJ/mol, while the model predicts ΔE_{pred}(D)= −18.93 kJ/mol. Example E and F show decreased distances of $d_{\text{I}\cdots\pi\text{-plane}}$(E) = 2.45 Å, and $d_{\text{I}\cdots\pi\text{-plane}}$(F) = 1.76 Å, respectively. Still observing high angle values, the decreased distance leads to repulsive interactions, with ΔE_{calc}(E) = 6.24 kJ/mol, and ΔE_{calc}(F) = 9.89 kJ/mol. In summary, all six outlier examples (A−F) lie outside the range of features represented in the training set, as reflected by their high MDs. Examples A−C exhibit geometries characteristic of close $\pi\cdots\pi$ interactions, which are not well represented in the training data. As a result, the model fails to properly capture these types of interaction and underestimates their energetic contributions. In contrast, examples D−F involve very close C−H$\cdots\pi$ contacts leading to repulsions that are not precisely accounted for in the model. Since such geometries were not explicitly included in the training set, the model naturally lacks the necessary information to recognize and appropriately penalize these interaction patterns. However, the model's performance on repulsive geometries is not crucial, as such interactions are typically identified and filtered out during earlier stages of the scoring process or by separate repulsive terms. Still, incorporating representative repulsive configurations of σ-hole interactions in future training could further improve the model's completeness and transferability. The same applies to recognition of $\pi\cdots\pi$ interactions. The model will be explicitly used for scoring halogen$\cdots\pi$ interactions with respect to the σ-hole interacting with aromatic moieties.

PDB Scan for Halogen$\cdots\pi$ Interactions in Crystal Structures. To apply the model to real-world biological examples, a PDB scan was conducted where we focused on protein−ligand recognition from a drug discovery perspective, thus excluding all types of halogenated biomolecular building blocks. 239,149 crystal structures (as of July 2025) were analyzed with 9810 (4.1%) unique structures containing ligands that bear chlorine, bromine, or iodine connected to an aromatic moiety. Initially, halogen$\cdots\pi$ contacts were considered if the aromatic side chain residue of phenylalanine, tyrosine, histidine, or tryptophan were located within 5 Å of the halogen atom. Results of the PDB scan are summarized in Table 2. A total of 23,536 contacts in 6796 unique PDB structures were found.

Since the model is supposed to assess halogen interactions onto benzene, we only focus on phenylalanine. With 10,174 (43.23%) contacts, halogen$\cdots$PHE contacts are the most prevalent. To concentrate on potential halogen bonds in terms of σ-hole interactions, we applied distance and angle constraints. Furthermore, the C−X vector of the ligand must point toward the benzene plane to increase the likelihood of a σ-hole interaction. Based on these restrictions, we identified 1114 (10.95% from the initial total PHE contacts) XB interactions for subsequent analysis. In a matched molecular pair approach, ligands were replaced with a model system of the corresponding halobenzene. The halogen atom and the C−X vector are matched exactly onto the original ligand. Additionally, the plane of the benzene ring is optimally aligned onto the ligand's aromatic ring system. The benzene ring of phenylalanine side chain residue was capped at the C_γ−C_β-bond and replaced by a protonated and optimized benzene system by aligning the two ring systems. It should be noted that the halobenzene model system can be much less tuned than original ligand systems where thewere calculated. Two examples were excluded, resulting scaffold or functional groups affect the σ-hole strength. The final data set consists of 1114 unique interaction geometries with 806 chlorine, 198 bromine, and 110 iodine interactions. Single-point calculations on an MP2/TZVPP level of theory were conducted using TURBOMOLE.

Table 2. PDB Scan Results for Halogen$\cdots\pi$ Contacts and Halogen$\cdots\pi$ Interactions[a]

addressed AA	halogen$\cdots\pi$ contacts within 5 Å (% of total)	halogen$\cdots\pi$ interaction after applied filters (% of contacts)
histidine	3597 (15.28%)	661 (18.38%)
phenylalanine	**10,174 (43.23%)**	**1114 (10.95%)**
tryptophan	2855 (12.13%)	384 (13.45%)
tyrosine	6910 (29.36%)	1152 (16.67%)
total	23,536	3311 (14.07%)
Phenylalanine Interactions per Halogen		
Cl	7667 (75.36%)	806 (10.51%)
Br	1874 (18.42%)	198 (10.57%)
I	633 (6.22%)	110 (17.38%)

[a]The focus is on phenylalanine contacts and interactions with corresponding results reported for each halogen separately.

Evaluation of the Model Using the PDB Set. Similar to the previous data set, adduct formation energies of the PDB set in a final set of 1112 complexes. These examples originate from the crystal structures 4HSX and 7EPZ, containing two geometries with very low distances facilitating severe clashes, and in turn, leading to repulsive energies. Geometric features were extracted from the geometries and fed into the model to assess its applicability to real-world biological examples, rather than relying solely on artificially derived data. The individual data points and energies are provided in spreadsheet format in the Supporting Information. As shown in Figure 5a, the distribution of energy differences ($\Delta\Delta E$) between calculated and predicted interaction energies remains centered around zero for all halogens (Cl, Br, I), with most differences falling within ±1 kJ/mol (82.22%). A maximum of $\Delta\Delta E$ = +20.42 kJ/mol and a minimum of $\Delta\Delta E$ = −6.49 kJ/mol was observed. The model achieves good predictive accuracy on the PDB examples, with an R^2 = 0.805 and an RMSE = 1.57 kJ/mol (Figure 5b). Despite the overall good accuracy, a few outliers (labeled A−D) still show notable prediction errors. These structures correspond to interaction geometries that lie statistically at the edge of, or outside the training feature distribution, as indicated by increased MDs. Detailed information on the four examples is given in Table 3. Figure 6 illustrates the detailed geometries of these four outlier examples (A−D) identified in the PDB data set analysis (Figure 5b). These structures show substantial deviations between predicted and calculated interaction energies. Example A displays a close contact with $d_{\text{Br}\cdots\pi\text{-plane}}$ = 1.68 Å and a relatively large angle of $\alpha_{\text{C-Br}\cdots\perp(\pi\text{-plane})}$ = 43.2°. The halogen, and particularly the C−X bond vector, is oriented

https://doi.org/10.1021/acs.jcim.5c02136
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

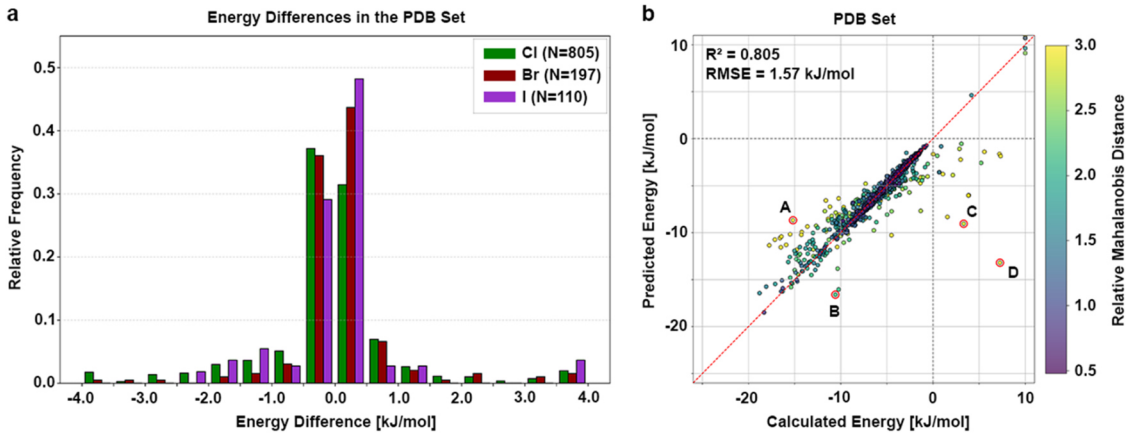


Figure 5. Model performance on the PDB set. (a) The histogram shows the relative frequencies of energy differences between calculated and predicted energy (total of 1112 data points) in bins of 0.5 kJ/mol from −4.0 to 4.0 kJ/mol for chlorine (green, N = 805), bromine (dark red, N = 197), and iodine (purple, N = 110) separately. Energy differences are calculated as $\Delta\Delta E = \Delta E_{calc} - \Delta E_{pred}$. Larger values are clipped to the respective limitation for better visibility. (b) Calculated energy is plotted against the predicted energy. The model achieved an R^2 = 0.805 with an RMSE = 1.57 kJ/mol. The red, dashed line indicates the perfect correlation between calculated and predicted values, while the gray dashed lines indicate the transition from negative to positive energies. Each data point is colored according to its relative MD with respect to the given color scale. Data points outlined with a red circle and labeled with (A−D) are discussed in detail in Figure 6.

Table 3. Overview of Energy Values and Geometric Parameters for Outlier Structures (A−D) (from the PDB Set) Highlighted in Figure 5 and Depicted in Figure 6[a]

	halogen	PDB ID	ΔE_{calc} [kJ/mol]	ΔE_{pred} [kJ/mol]	$\Delta\Delta E$ [kJ/mol]	$d_{X\cdots\pi\text{-plane}}$ [Å]	$\alpha_{C-X\cdots\perp(\pi\text{-plane})}$ [deg]
A	Br	5RTQ	−15.17	−8.68	−6.49	1.68	43.2
B	Cl	5CU3	−10.58	−16.6	6.02	2.15	44.6
C	Cl	2YLP	3.33	−9.06	12.39	1.61	40.1
D	Cl	2Q6N	7.26	−13.2	20.47	1.25	38.3

[a]The halogen symbolizes the interacting halobenzene. The PDB ID indicates the crystal structure the interaction geometry was extracted from. Values of calculated and predicted energies, as well as the difference between both are given in kJ/mol. Distance values $d_{X\cdots\pi\text{-plane}}$ between the halogen and the π-plane are given in Å. Angle values between the C−X vector and the normal to the π-plane are given in degree [deg].

away from the aromatic plane. This orientation places the σ-hole outside the optimal interaction direction. However, this geometry exhibits features of other interaction types, resulting in a mixed interaction profile that combines aspects of $\pi\cdots\pi$ stacking and lateral C−H···X contacts onto the negative belt of the halogen. The overall interaction remains attractive, but its nature differs from the purely σ-hole-driven contacts represented in the training data. Therefore, this configuration lies outside the range of training data, leading to a notable underestimation by the model. In contrast, examples B, C, and D are characterized by short halogen−π distances ($d_{Cl\cdots\pi\text{-plane}}$ = 2.15 Å, 1.61 Å, and 1.25 Å, respectively) and moderate angles ($\alpha_{C-Br\cdots\perp(\pi\text{-plane})}$ = 44.6°, 40.1°, and 38.3°), yet the model overestimates their interaction strength. These discrepancies likely arise from the fact that these geometries, although derived from biological examples, are underrepresented or absent in the training distribution, limiting the model's ability to assign accurate energies. Importantly, these interaction motifs fall outside the intended scope of σ-hole interactions. In these cases, the halogen atom is positioned outside the π-system plane, or the halobenzene adopts a geometry with unrealistically short contact distances to the benzene molecule, leading to steric repulsion as confirmed by positive ΔE_{calc} values. Additionally, the C−X bond vector typically points away from the π-system, preventing the spatial arrangement required for a directional σ-hole interaction. The origin of these apparent steric clashes in examples B (2YLP) and C (2Q6N) is further examined in the Supporting Information, where we provide a more detailed structural analysis (Figures S2 and S3). Thus, while such geometries may appear in experimental structures, they do not constitute σ-hole interactions and are neither expected nor intended to be well captured or prioritized by the model. In addition, although the model was trained exclusively on halobenzenes, some PDB examples contain halogens bound to heteroaromatic five-membered rings. To evaluate transferability, we recalculated the interaction features using the corresponding five-ring geometries and compared the predicted energies to those obtained with the initial six-membered ring features of the halobenzene. On average, the difference between both predictions was 0.3 kJ/mol with a standard deviation of 0.44 kJ/mol, demonstrating that the model generalizes well to heteroaromatic π-systems and that the feature representation is not biased toward a particular ring type.

For practical applications such as docking algorithms or virtual screening, additional empirical terms for $\pi\cdots\pi$ or C−H···π interactions would be required to properly capture these motifs. Addressing such effects lies beyond the present scope of this study. However, repulsive geometric arrangements are typically already handled by standard distance-dependent terms in scoring functions, while attractive motifs such as $\pi\cdots\pi$ or C−H···π interactions should or could be incorporated through dedicated scoring functions or extensions of existing ones.

Conclusion and Outlook. In this proof-of-concept study, we have employed neural networks (NNs) to predict the

https://doi.org/10.1021/acs.jcim.5c02136
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

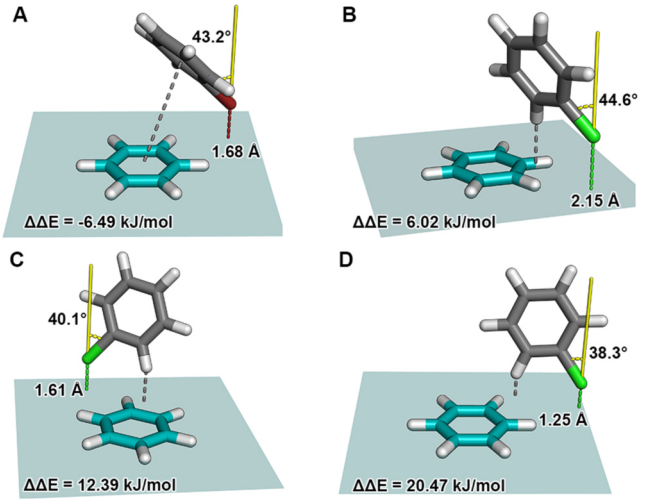


Figure 6. Interaction geometries corresponding to selected outliers of the PDB set identified in the scatter plot shown in Figure 5. These structures represent data points with unusually high deviations between calculated and predicted adduct formation energies. Each geometry (A–D) illustrates the spatial arrangement of the halobenzene (gray) relative to the π-system of benzene (teal), including the halogen distance $d_{X\cdots\pi\text{-plane}}$ (dashed line, colored according to the halogen) to the π-system plane in Å, the torsion angle $\alpha_{C-X\cdots\perp(\pi\text{-plane})}$ (colored yellow) in degrees between the C–X vector and the normal of the π-system plane, as well as the energy difference $\Delta\Delta E$. The gray dashed line highlights the type of interaction that may contribute to the observed prediction error. For better visibility, the teal-colored plane illustrates the benzene plane and the dimensions of the models' training grid. (A) Bromobenzene interaction with $d_{Br\cdots\pi\text{-plane}}$ = 1.68 Å and $\alpha_{C-Br\cdots\perp(\pi\text{-plane})}$ = 43.2° The gray dashed line indicates a $\pi\cdots\pi$ interaction. (B) Chlorobenzene interaction with $d_{Cl\cdots\pi\text{-plane}}$ = 2.15 Å and $\alpha_{C-Cl\cdots\perp(\pi\text{-plane})}$ = 44.7°. The gray dashed line indicates a C–H$\cdots\pi$ contact. (C) Chlorobenzene interaction with $d_{Cl\cdots\pi\text{-plane}}$ = 1.61 Å and $\alpha_{C-Cl\cdots\perp(\pi\text{-plane})}$ = 40.1°. The gray dashed line indicates a C–H$\cdots\pi$ contact. (D) Chlorobenzene interaction with $d_{Cl\cdots\pi\text{-plane}}$ = 1.25 Å and $\alpha_{C-Cl\cdots\perp(\pi\text{-plane})}$ = 38.3°. The gray dashed line indicates a C–H$\cdots\pi$ contact.

interaction strength of halogen$\cdots\pi$ interactions between halobenzenes (chlorobenzene, bromobenzene, iodobenzene) and a simple benzene model. Nearly 1.4 million single-point calculations were conducted and adduct formation energies calculated. NN models were trained on geometric features extracted from the interaction geometries. Extensive hyperparameter tuning was used to find the most suitable model configuration. Model validation was carried out using stratified 5-fold cross-validation and further tested on independent data sets. The model demonstrated excellent predictive performance on the cross-validation set, achieving an R^2 = 0.998 and a very low RMSE = 0.16 kJ/mol, a level of accuracy that likely approaches the intrinsic limits of computability of interaction energies. These results reflect the model's ability to interpolate accurately within the feature space of the training data. When applied to a separate random geometry test set, the model still performed strongly with R^2 = 0.86 and RMSE = 1.33 kJ/mol, while similar results were obtained for the PDB-derived data set with R^2 = 0.81 and RMSE = 1.57 kJ/mol. Most predictions fall within a narrow error range, demonstrating the model's ability to generalize across a wide variety of geometries. Analysis of outliers revealed that large prediction errors occur primarily when input features lie far beyond the training distribution. In such cases, the model either fails to recognize favorable $\pi\cdots\pi$ interaction motifs or misinterprets repulsive C–H$\cdots\pi$ contacts, which were not explicitly covered during training. Nonetheless, these geometries are not representative of the directional σ-hole interactions the model was designed to capture and can therefore be neglected or are covered by other terms. Nevertheless, incorporating the repulsive part of σ-hole interactions in future developments would improve the robustness of the model and extend its applicability beyond the attractive interaction space. Overall, this study demonstrates that NNs trained on well-defined and well-chosen geometric features can reliably predict halogen$\cdots\pi$ interaction energies within their defined purpose. The approach offers a fast and accurate alternative to quantum chemical calculations. By leveraging the methodological "double jump" from CCSD(T) → MP2 → NNs, the model retains accuracy close to CCSD(T) benchmarks while achieving a runtime speed-up of up to 8 orders of magnitude compared to MP2 calculations. However, this study represents an initial step involving only a single aromatic residue model. As a next step, we aim to extend this approach to the remaining aromatic amino acid side chain residues of histidine, tyrosine, and tryptophan, for which data generation and model development are already underway. By constructing exclusive NN models for each of these residues, we seek to capture the diversity of biologically relevant halogen$\cdots\pi$ interactions. Ultimately, these models will be integrated into the molecular docking framework PLANTS, enabling improved recognition and scoring of halogen bonding interactions in protein–ligand complexes. Additionally, we aim to gradually integrate the models within the Galaxy Webserver[58] as standalone scoring function modules. Uploaded molecular structures will be systematically analyzed to detect halogen$\cdots\pi$ interactions, and, if present, the interaction strengths will be quantitatively evaluated.

MATERIALS AND METHODS

Structure Optimization. Geometry optimizations of the individual ligand model systems (iodobenzene, bromobenzene, and chlorobenzene) and the amino acid model system (benzene) were done at the MP2[59,60]-level of theory using TURBOMOLE 7.7.1[61] with a triple-ζ basis set (def2-TZVPP[62]) on the JUSTUS2–bwHPC Cluster.[63] Calculations were done in combination with the resolution of identity (RI) technique and the frozen core approximation. Frozen core orbitals were defined using default settings, where orbitals with energies below −3.0 au are considered core orbitals. SCF convergence criterion was increased to 10^{-8} hartree. Relativistic effects for iodine were considered by an effective core potential (ECP).[64−72]

Generation of Interaction Geometries. Interaction geometries of chloro-, bromo-, and iodobenzene in complex with benzene were generated. Halobenzenes were placed on a regular grid using *X*- and *Z*-translations (Figure 7a) for eight different distances (Figure 7b). Following previous approaches, an optimal σ-hole angle of $\alpha_{C\text{-}X\cdots\pi\text{-plane}}$ = 180° was initially used. To capture rotational features, the halobenzene itself was rotated around the *y*-axis. Furthermore, the σ-hole angle was altered to deviations from −40° to 40° in steps of 10° in the *x*-, and *z*-directions (Figure 7d). Two additional rotation axes were incorporated, lying in 45° to the *x*-, or *z*-axis (Figure 7e). Due to the symmetric nature of benzene, only one quadrant of the grid was considered. With this systematic procedure, 508,032 interaction geometries per halobenzene were

https://doi.org/10.1021/acs.jcim.5c02136
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

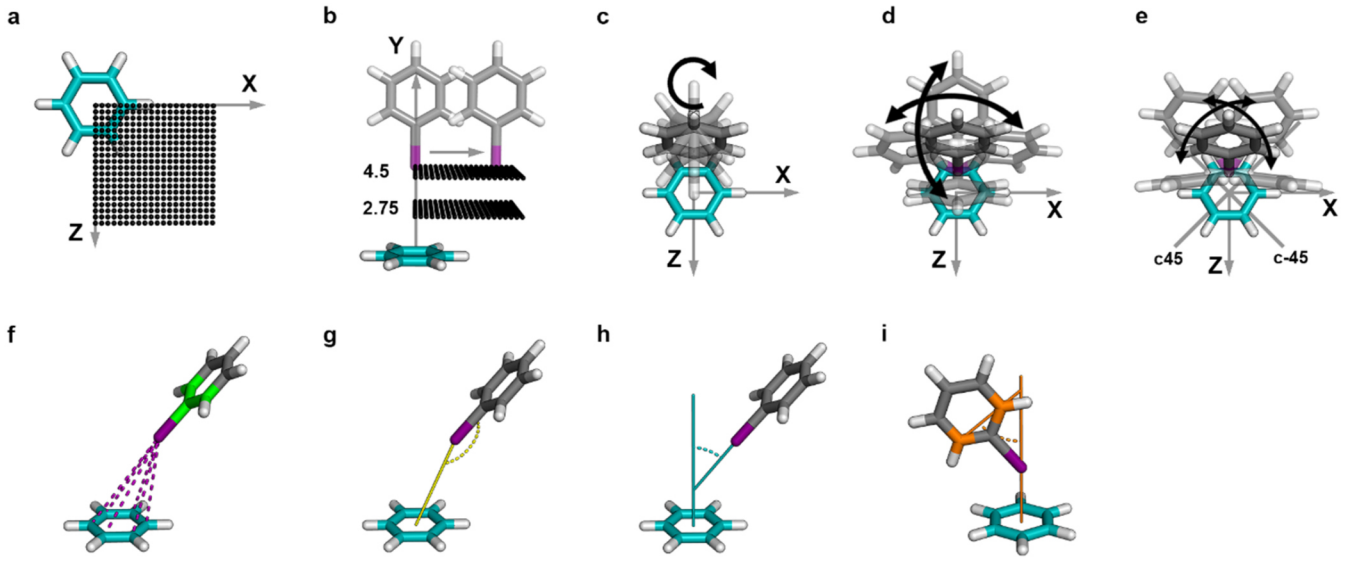


Figure 7. Overview of the interaction geometry generation and the feature extraction. (a) Grid points on the *XZ*-plane were generated with dimensions $X_{translation}$ = [0.0 Å to 5.0 Å], $Z_{translation}$ = [0.0 Å to 5.0 Å] in steps of 0.25 Å. (b) Grid points were generated for eight different distances, $d_{X\cdots\pi\text{-plane}}$ = [2.75 Å to 4.5 Å] in steps of 0.25 Å, between the halogen atom (Cl, Br, or I) and the benzene plane. (c) Rotations of the halobenzene around the *y*-axis y_{rot} = [0° (initially), 45°, 90°, 135°]. (d) Deviations from the optimal σ-hole angle $\alpha_{C-X\cdots\pi\text{-plane}}$ = 180° from −40° to 40° in steps of 10° achieved by rotating around the *x*- and *z*-axis. (e) Custom-generated rotational axis (c45 and c-45), lying in 45° to the *x*-, and *z*-axis. Rotations around these axes similar to (d). (f) Pairwise distances from the halogen atom and the green colored carbons of the halobenzene to all carbon atoms of the benzene are extracted as features. (g) Angle feature of the halogen atom, its neighboring carbon and the center of mass of the benzene, $\alpha_{C-X\cdots CoM(benzene)}$. (h) Angle feature of the C−X vector and the normal of the benzene plane $\alpha_{C-X\cdots\perp(\pi\text{-plane})}$. (i) Angle feature of the vector between the orange-colored carbons of the halobenzene and the normal of the benzene plane.

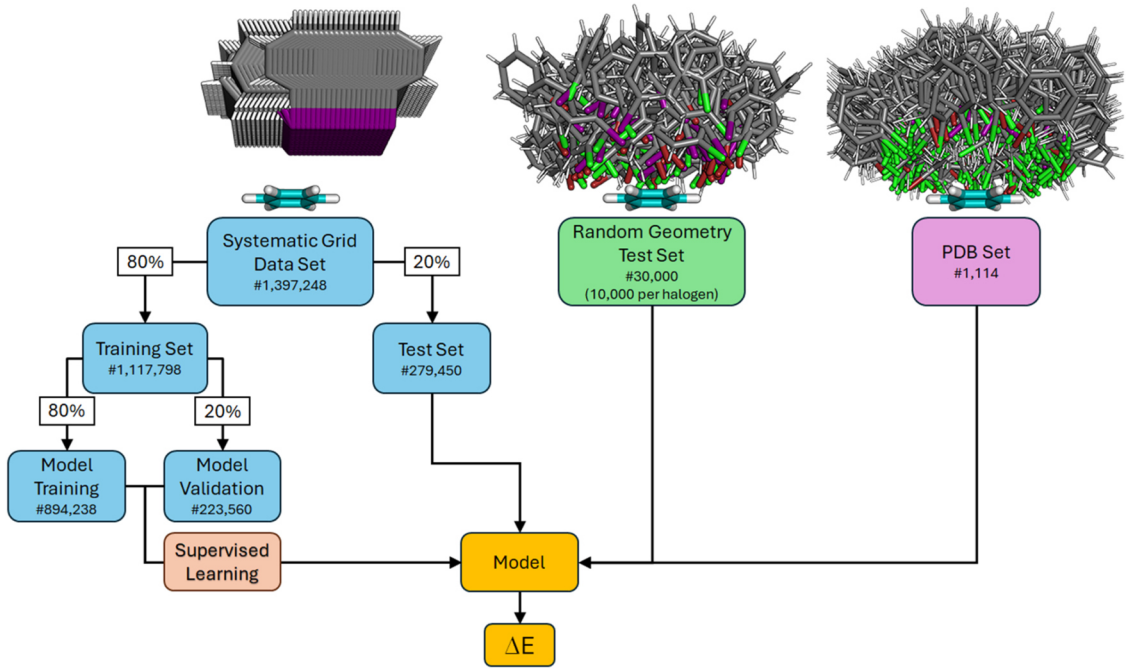


Figure 8. Overview of the different data sets. The systematic grid data set (1,397,248 geometries) is split into training (80% of the whole data set, 1,117,798 geometries) and test set (20%, 279,450 geometries). The training set is further split into model training (again 80%, 894,236 geometries) and model validation set (20%, 223,560 geometries). The two model data sets are fed into the model in a supervised training approach. The random geometry test set (30,000 geometries) is used to evaluate the models' generalized performance on unseen data. The PDB set (1114 geometries) is used to represent and evaluate biological examples. The respective geometries shown are only a small excerpt of the full data sets.

generated. A total of 1,397,247 single point calculations were conducted on an MP2/TZVPP level of theory. Adduct formation energies were calculated as

$$\Delta E_{calc} = (E_{complex} - (E_{halobenzene} + E_{benzene})) \quad (1)$$

and reported as kJ/mol.

Feature Extraction and NN Model Training. All data preprocessing, feature extraction, and learning approaches were built in Python using custom scripts with the *PyTorch*[73] and *scikit-learn* package, two open-source Python libraries for

https://doi.org/10.1021/acs.jcim.5c02136
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

machine learning. NN models were trained on geometric features extracted from the interaction geometries. The training process was performed on the BinAC−bwHPC Cluster.[74] The feature vector $\vec{\mathbf{v}}$= (d_1, d_2, ..., d_n, a_1, a_2, a_3) consists of 30 features. Table S1 shows details of all features. The adduct formation energy ΔE of each geometry serves as the target value. Distance features are calculated as pairwise distances between selected atoms of the halobenzene (halogen atom, its neighboring carbon atom, and from this carbon two adjacent carbon atoms) and the benzene system (Figure 7f). The features are sorted according to the distances between halogen and benzene carbon atoms in ascending order, such that the nearest distance is defined as $d_{X\cdots C}(\min) = \vec{\nu}(d_1)$, the second-shortest distance as $\vec{\nu}(d_2)$, and so forth. The distances between the benzene carbons and the halobenzene carbons follow the same ordering as the halogen−carbon distances. Three angle features are calculated between vectors of the halobenzene and the benzene plane. One angle is calculated between the halogen atom, its neighboring carbon atom, and the center of mass of the benzene (Figure 7g). A second angle is calculated between the C−X vector and the normal of the benzene plane (Figure 7h). And a third is calculated between two carbon atoms of the halobenzene and the normal of the benzene plane (Figure 7i). Before training, all features were individually normalized using a min-max scaler (scikit-learn MinMaxScaler). The data set was split into training (80% of the data) and test set (20%) in a stratified leave-one-out 5-fold cross validation approach to ensure model consistency (Figure 8). Stratification was applied using (1) the halogen, and (2) distances between halogen and π-plane to ensure features were equally distributed and divided across the corresponding sets. Subsequently, the training set was further split into model training (80%) and model validation (20%) using the same stratification strategy. Extensive hyperparameter tuning was performed on the training data set within a supervised learning approach. This process involved systematic variation of activation functions (Sigmoid, Tanh, and Leaky ReLU, as provided by *PyTorch*), the number of hidden layers (ranging from two to six), and the sizes of hidden layers using combinations from the set [256, 128, 64, 32, 16, 8] in a fully connected feed-forward network. Additional parameters included initial learning rates (10^{-1} to 10^{-5}), batch sizes (32, 64, 128, 256), and the number of training epochs (10 to 1000). Model training utilized the Adam optimizer (*PyTorch* built-in), an elastic-net-weighted-MSE loss function to address data imbalance, and early stopping criteria to prevent overfitting. The combination yielding the best performance on the validation set was selected for final training and evaluation. The mean absolute error (MAE), the root-mean-square error (RMSE), and the coefficient of determination (R^2) were employed to evaluate the predictive performance of the models. An R^2 value approaching 1.00, alongside low MAE and RMSE values, indicates that a model achieves high predictive accuracy. Energy differences between calculated and predicted adduct formation energies are reported as

$$\Delta\Delta E = \Delta E_{\text{calculated}} - \Delta E_{\text{predicted}} \quad (2)$$

in kJ/mol. The training started with a fixed learning rate and was adapted (value was halved) up to a minimum of 0.0001 if there had not been a loss improvement for 10 epochs. The final model consists of three internal hidden layers (of size: 64, 32, 16) with Leaky Rectified Linear Unit (Leaky ReLU) as activation functions, an initial learning rate of 0.01, and was trained on a batch size of 128.

Generation of a Random Geometry Test Set. Similar to the generation of the systematic data set used for the training, we generated a smaller subset of 30,000 interaction geometries (10,000 per halobenzene) with random geometric features to test the model's generalized performance on unseen data. Therefore, parameters of $X,Z_{\text{translation}}$ = [−5.0 Å to 5.0 Å], $Y_{\text{translation}}$ = [1.5 Å to 5.0 Å], y_{rot} = [0° to 360°], $\alpha_{C-X\cdots\perp(\pi\text{-plane})}$ = [0° to 60°] were randomly chosen and applied to a halobenzene. To ensure uniqueness of the generated geometries, newly chosen parameters were compared to previous ones, and only applied if found to be distinct.

Such a more inclusive definition of broad angular and distance limits was employed to capture a diverse range of halogen···π interaction geometries, including those near the characteristic boundaries of σ-hole interactions. Our aim is to evaluate the model's generalizability and highlight its potential limitations.

PDB Scan of Phenylalanine Acceptors. A PDB (as of July 2025:239,149) scan was conducted using a custom Python/PyMOL[75] script. Alternative conformations, metal ions and hydrogen atoms were removed from the structure. PDB structures containing halogenated ligands (chlorine, bromine, iodine) with the halogen atom connected to an aromatic ring system and a ligand size of six or more heavy atoms were considered for further analysis. Aromatic amino acid side chain residues of phenylalanine within 5 Å of the halogen atom were retained as XB acceptors. As an additional constraint, the C−X vector must point toward the benzene plane ($\alpha_{C-X\cdots\perp(\pi\text{-plane})} < 50°$) increasing the likelihood of a potential halogen bond. As lower distance limit we chose a minimal distance of $d_{X\cdots AA}$= 1 Å. The side chain residue was replaced by a benzene molecule. For simplicity, we replaced each ligand with the corresponding halobenzene (chloro-, bromo, or iodobenzene) by matching the halogen atom, the C−X bond vector and the plane of the (hetero)aromatic system in a matched molecular pair approach. Both molecules were separately geometry optimized on an MP2/TZVPP level of theory as previously described. A total of 1114 interaction geometries of the halobenzene in complex with the addressed benzene were extracted to carry out a single point calculation on an MP2/TZVPP level of theory. They again represent a more inclusive set of diverse geometries, even at the border of typical halogen bonds. Adduct formation energies were calculated using the supermolecular approach.

Model Evaluation and Outlier Detection. To evaluate the model's ability to predict adduct formation energies of a given interaction geometry we use the three data sets (20% test set, random geometry set, and PDB set). Ideally, the data sets consist of "unseen" data for the model to test its' generalizability. Similar to the training process, we examined the results of the prediction using the root-mean-square error (RMSE) and the coefficient of determination (R^2). Given that interaction geometries in the random geometry set and the PDB set can lie outside the feature space of the training set, the Mahalanobis Distance[76] (MD), a statistical measure of dissimilarity between each test point and the center of the multivariate distribution of training features, was employed to assess the degree to which a given data point deviates from the input space. It is defined as

$$D_M(x) = \sqrt{(x-\mu)^T \Sigma^{-1} (x-\mu)} \quad (3)$$

https://doi.org/10.1021/acs.jcim.5c02136
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

where x is the feature vector of the data point, μ is the mean vector of the training data, and $\sum$ is the covariance matrix of the training data. High distances indicate that a geometry lies in a statistically rare region of input space relative to the training set, thus identifying it as a potential extrapolation point or outlier. Relative MDs were computed for all geometries in the test sets, and a threshold for outlier classification was set based on the 95th percentile of the distribution of distances within the training set.

■ ASSOCIATED CONTENT

Data Availability Statement

PyMOL is an open-source software maintained and distributed by Schrödinger. There is an open-source version of PyMOL available at: https://github.com/schrodinger/pymol-open-source. Python and all of its' packages is an open-source programming language available and downloadable from https://www.python.org/. PyTorch is an open-source machine learning library for Python: https://pytorch.org/. TURBOMOLE is a purchasable software maintained and distributed by the *TURBOMOLE GmbH*. Demo versions are available at https://www.turbomole.org/. The licensed software was provided to us by the bwHPC Cluster JUSTUS2.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.5c02136.

Graphical depiction of the distance-based features; Detailed description of the features used for training; Structural analysis of PDB examples 2YLP, 26QN (PDF)

Detailed tables of the individual data points of the random geometry data set and the PDB data set with corresponding adduct formation energies for calculated and predicted energies (XLSX)

■ AUTHOR INFORMATION

Corresponding Author

Frank M. Boeckler − *Laboratory for Molecular Design & Pharmaceutical Biophysics, Institute of Pharmaceutical Sciences, Department of Pharmacy and Biochemistry, Eberhard Karls Universität Tübingen, 72076 Tübingen, Germany; Interfaculty Institute for Biomedical Informatics (IBMI), Eberhard Karls Universität Tübingen, 72076 Tübingen, Germany;* orcid.org/0000-0001-8738-6716; Email: frank.boeckler@uni-tuebingen.de

Authors

Marc U. Engelhardt − *Laboratory for Molecular Design & Pharmaceutical Biophysics, Institute of Pharmaceutical Sciences, Department of Pharmacy and Biochemistry, Eberhard Karls Universität Tübingen, 72076 Tübingen, Germany;* orcid.org/0009-0007-9152-8538

Finn Mier − *Laboratory for Molecular Design & Pharmaceutical Biophysics, Institute of Pharmaceutical Sciences, Department of Pharmacy and Biochemistry, Eberhard Karls Universität Tübingen, 72076 Tübingen, Germany*

Markus O. Zimmermann − *Laboratory for Molecular Design & Pharmaceutical Biophysics, Institute of Pharmaceutical Sciences, Department of Pharmacy and Biochemistry, Eberhard Karls Universität Tübingen, 72076 Tübingen, Germany; Interfaculty Institute for Biomedical Informatics (IBMI), Eberhard Karls Universität Tübingen, 72076 Tübingen, Germany;* orcid.org/0000-0001-6115-8248

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.5c02136

Author Contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript. F.M.B. and M.U.E. envisioned the research. M.U.E. performed all QM calculations, developed all scripts for machine learning and evaluation, gathered all results and prepared the corresponding visualizations. M.U.E. prepared the original draft. F.M. contributed to developing the computational strategy and provided comments on the manuscript. M.U.E., M.O.Z., F.M., and F.M.B. reviewed, edited, and finalized the manuscript.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors acknowledge support from the state of Baden−Württemberg through bwHPC and the German Research Foundation (DFG) through Grant No. INST 40/575-1 FUGG (JUSTUS 2 cluster). In addition, support is acknowledged from the High Performance and Cloud Computing Group at the Zentrum für Datenverarbeitung of the University of Tübingen and the German Research Foundation (DFG) through Grant No. INST 37/935-1 FUGG (BinAC cluster).

■ ABBREVIATIONS

XB, halogen bond; X, halogen; QM, quantum mechanical; MP2, Møller−Plesset perturbation method order 2; TZVPP, valence triple-ζ with two sets of polarization functions; cc-pVTZ, correlation-consistent polarized valence triple-ζ function; SCF, self-consistent field; RI, resolution of identity; RMSE, root mean square error; ECP, effective core potential; NN, neural network; R^2, coefficient of determination; MD, Mahalanobis Distance; PDB, Protein Data Bank; PHE, phenylalanine; MAE, mean absolute error; MSE, mean squared error; CCSD(T), coupled cluster with single, double, and perturbative triple excitations

■ REFERENCES

(1) Bissantz, C.; Kuhn, B.; Stahl, M. A Medicinal Chemist's Guide to Molecular Interactions. *J. Med. Chem.* **2010**, *53* (14), 5061−5084.

(2) Müller-Dethlefs, K.; Hobza, P. Noncovalent Interactions: A Challenge for Experiment and Theory. *Chem. Rev.* **2000**, *100* (1), 143−168.

(3) Anighoro, A. Underappreciated Chemical Interactions in Protein−Ligand Complexes. In *Quantum Mechanics in Drug Discovery*; Heifetz, A., Ed.; Springer US, 2020; pp 75−86.

(4) Adhav, V. A.; Saikrishnan, K. The Realm of Unconventional Noncovalent Interactions in Proteins: Their Significance in Structure and Function. *ACS Omega* **2023**, *8* (25), 22268−22284.

(5) Jena, S.; Dutta, J.; Tulsiyan, K. D.; Sahu, A. K.; Choudhury, S. S.; Biswal, H. S. Noncovalent interactions in proteins and nucleic acids: beyond hydrogen bonding and π-stacking. *Chem. Soc. Rev.* **2022**, *51* (11), 4261−4286.

(6) Cavallo, G.; Metrangolo, P.; Milani, R.; Pilati, T.; Priimagi, A.; Resnati, G.; Terraneo, G. The Halogen Bond. *Chem. Rev.* **2016**, *116* (4), 2478−2601.

(7) Clark, T.; Hennemann, M.; Murray, J. S.; Politzer, P. Halogen bonding: the σ-hole. *J. Mol. Model.* **2007**, *13* (2), 291−296.

(8) Desiraju, G. R.; Ho, P. S.; Kloo, L.; Legon, A. C.; Marquardt, R.; Metrangolo, P.; Politzer, P.; Resnati, G.; Rissanen, K. Definition of the halogen bond (IUPAC Recommendations 2013). *Pure Appl. Chem.* **2013**, *85* (8), 1711−1713.
(9) Politzer, P.; Murray, J. S.; Clark, T. Halogen bonding: an electrostatically-driven highly directional noncovalent interaction. *Phys. Chem. Chem. Phys.* **2010**, *12* (28), 7748−7757.
(10) Wang, C.; Danovich, D.; Mo, Y.; Shaik, S. On The Nature of the Halogen Bond. *J. Chem. Theory Comput.* **2014**, *10* (9), 3726−3737.
(11) Erdélyi, M. Halogen bonding in solution. *Chem. Soc. Rev.* **2012**, *41* (9), 3547−3557.
(12) Heidrich, J.; Exner, T. E.; Boeckler, F. M. Predicting the Magnitude of σ-Holes Using VmaxPred, a Fast and Efficient Tool Supporting the Application of Halogen Bonds in Drug Discovery. *J. Chem. Inf. Model.* **2019**, *59* (2), 636−643.
(13) Bhattarai, S.; Sutradhar, D.; Chandra, A. K. Tuning of halogen-bond strength: Comparative role of basicity and strength of σ-hole. *J. Mol. Struct.* **2021**, *1223*, No. 129239.
(14) Donald, K. J.; Pham, N.; Ravichandran, P. Sigma Hole Potentials as Tools: Quantifying and Partitioning Substituent Effects. *J. Phys. Chem. A* **2023**, *127* (48), 10147−10158.
(15) Esrafili, M. D.; Mahdavinia, G.; Javaheri, M.; Sobhi, H. R. A theoretical study of substitution effects on halogen−π interactions. *Mol. Phys.* **2014**, *112* (8), 1160−1166.
(16) Andrea, R. V.; P, S. H. The Role of Halogen Bonding in Inhibitor Recognition and Binding by Protein Kinases. *Curr. Top. Med. Chem.* **2007**, *7* (14), 1336−1348.
(17) Auffinger, P.; Hays, F. A.; Westhof, E.; Ho, P. S. Halogen bonds in biological molecules. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101* (48), 16789−16794.
(18) Berger, G.; Frangville, P.; Meyer, F. Halogen bonding for molecular recognition: new developments in materials and biological sciences. *Chem. Commun.* **2020**, *56* (37), 4970−4981.
(19) Hardegger, L. A.; Kuhn, B.; Spinnler, B.; Anselm, L.; Ecabert, R.; Stihle, M.; Gsell, B.; Thoma, R.; Diez, J.; Benz, J.; et al. Systematic Investigation of Halogen Bonding in Protein−Ligand Interactions. *Angew. Chem., Int. Ed.* **2011**, *50* (1), 314−318.
(20) Jiang, L.; Zhang, X.; Zhou, Y.; Chen, Y.; Luo, Z.; Li, J.; Yuan, C.; Huang, M. Halogen bonding for the design of inhibitors by targeting the S1 pocket of serine proteases. *RSC Adv.* **2018**, *8* (49), 28189−28197.
(21) Parisini, E.; Metrangolo, P.; Pilati, T.; Resnati, G.; Terraneo, G. Halogen bonding in halocarbon−protein complexes: a structural survey. *Chem. Soc. Rev.* **2011**, *40* (5), 2267−2278.
(22) Parker, A. J.; Stewart, J.; Donald, K. J.; Parish, C. A. Halogen Bonding in DNA Base Pairs. *J. Am. Chem. Soc.* **2012**, *134* (11), 5165−5172.
(23) Voth, A. R.; Khuu, P.; Oishi, K.; Ho, P. S. Halogen bonds as orthogonal molecular interactions to hydrogen bonds. *Nat. Chem.* **2009**, *1* (1), 74−79.
(24) Xu, Z.; Yang, Z.; Liu, Y.; Lu, Y.; Chen, K.; Zhu, W. Halogen Bond: Its Role beyond Drug−Target Binding Affinity for Drug Discovery and Development. *J. Chem. Inf. Model.* **2014**, *54* (1), 69−78.
(25) Dammann, M.; Stahlecker, J.; Zimmermann, M. O.; Klett, T.; Rotzinger, K.; Kramer, M.; Coles, M.; Stehle, T.; Boeckler, F. M. Screening of a Halogen-Enriched Fragment Library Leads to Unconventional Binding Modes. *J. Med. Chem.* **2022**, *65* (21), 14539−14552.
(26) Politzer, P.; Murray, J. S.; Clark, T. Halogen bonding and other σ-hole interactions: a perspective. *Phys. Chem. Chem. Phys.* **2013**, *15* (27), 11178−11189.
(27) Vaas, S.; Zimmermann, M. O.; Schollmeyer, D.; Stahlecker, J.; Engelhardt, M. U.; Rheinganz, J.; Drotleff, B.; Olfert, M.; Lämmerhofer, M.; Kramer, M.; et al. Principles and Applications of CF2X Moieties as Unconventional Halogen Bond Donors in Medicinal Chemistry, Chemical Biology, and Drug Discovery. *J. Med. Chem.* **2023**, *66* (15), 10202−10225.
(28) Wilcken, R.; Liu, X.; Zimmermann, M. O.; Rutherford, T. J.; Fersht, A. R.; Joerger, A. C.; Boeckler, F. M. Halogen-Enriched Fragment Libraries as Leads for Drug Rescue of Mutant p53. *J. Am. Chem. Soc.* **2012**, *134* (15), 6810−6818.
(29) Mitra, D.; Bankoti, N.; Michael, D.; Sekar, K.; Row, T. N. G. C-halogen···pi interactions in nucleic acids: a database study. *J. Chem. Sci.* **2020**, *132* (1), No. 93.
(30) Riley, K. E.; Ford, C. L.; Demouchet, K. Comparison of hydrogen bonds, halogen bonds, CH···π interactions, and CX···π interactions using high-level ab initio methods. *Chem. Phys. Lett.* **2015**, *621*, 165−170.
(31) Metrangolo, P.; Resnati, G. *Halogen Bonding: Fundamentals and Applications*; Springer, 2008 DOI: 10.1007/978-3-540-74330-9.
(32) Metrangolo, P.; Resnati, G. *Halogen Bonding I: Impact on Materials Chemistry and Life Sciences*; Springer, 2015 DOI: 10.1007/978-3-319-14057-5.
(33) Metrangolo, P.; Resnati, G. *Halogen Bonding II: Impact on Materials Chemistry and Life Sciences*; Springer, 2015 DOI: 10.1007/978-3-319-15732-0.
(34) Wilcken, R.; Zimmermann, M. O.; Lange, A.; Zahn, S.; Boeckler, F. M. Using halogen bonds to address the protein backbone: a systematic evaluation. *J. Comput. Aided Mol. Des.* **2012**, *26* (8), 935−945.
(35) Zimmermann, M. O.; Boeckler, F. M. Targeting the protein backbone with aryl halides: systematic comparison of halogen bonding and π···π interactions using N-methylacetamide. *MedChemComm* **2016**, *7* (3), 500−505.
(36) Wilcken, R.; Zimmermann, M. O.; Lange, A.; Zahn, S.; Kirchner, B.; Boeckler, F. M. Addressing Methionine in Molecular Design through Directed Sulfur−Halogen Bonds. *J. Chem. Theory Comput.* **2011**, *7* (7), 2307−2315.
(37) Lange, A.; Zimmermann, M. O.; Wilcken, R.; Zahn, S.; Boeckler, F. M. Targeting Histidine Side Chains in Molecular Design through Nitrogen−Halogen Bonds. *J. Chem. Inf. Model.* **2013**, *53* (12), 3178−3189.
(38) Zimmermann, M. O.; Lange, A.; Zahn, S.; Exner, T. E.; Boeckler, F. M. Using Surface Scans for the Evaluation of Halogen Bonds toward the Side Chains of Aspartate, Asparagine, Glutamate, and Glutamine. *J. Chem. Inf. Model.* **2016**, *56* (7), 1373−1383.
(39) Engelhardt, M. U.; Zimmermann, M. O.; Dammann, M.; Stahlecker, J.; Poso, A.; Kronenberger, T.; Kunick, C.; Stehle, T.; Boeckler, F. M. Halogen Bonding on Water—A Drop in the Ocean? *J. Chem. Theory Comput.* **2024**, *20* (23), 10716−10730.
(40) Zhou, P.; Lv, J.; Zou, J.; Tian, F.; Shang, Z. Halogen−water−hydrogen bridges in biomolecules. *J. Struct. Biol.* **2010**, *169* (2), 172−182.
(41) Heroven, C.; Georgi, V.; Ganotra, G. K.; Brennan, P.; Wolfreys, F.; Wade, R. C.; Fernández-Montalván, A. E.; Chaikuad, A.; Knapp, S. Halogen−Aromatic π Interactions Modulate Inhibitor Residence Times. *Angew. Chem., Int. Ed.* **2018**, *57* (24), 7220−7224.
(42) Matter, H.; Nazaré, M.; Güssregen, S.; Will, D. W.; Schreuder, H.; Bauer, A.; Urmann, M.; Ritter, K.; Wagner, M.; Wehner, V. Evidence for C-Cl/C-Br···π Interactions as an Important Contribution to Protein−Ligand Binding Affinity. *Angew. Chem., Int. Ed.* **2009**, *48* (16), 2911−2916.
(43) Shah, M. B.; Liu, J.; Zhang, Q.; Stout, C. D.; Halpert, J. R. Halogen−π Interactions in the Cytochrome P450 Active Site: Structural Insights into Human CYP2B6 Substrate Selectivity. *ACS Chem. Biol.* **2017**, *12* (5), 1204−1210.
(44) Saraogi, I.; Vijay, V. G.; Das, S.; Sekar, K.; Row, T. N. G. C−halogen···π interactions in proteins: a database study. *Cryst. Eng.* **2003**, *6* (2), 69−77.
(45) Wallnoefer, H. G.; Fox, T.; Liedl, K. R.; Tautermann, C. S. Dispersion dominated halogen−π interactions: energies and locations of minima. *Phys. Chem. Chem. Phys.* **2010**, *12* (45), 14941−14949.
(46) Engelhardt, M. U.; Zimmermann, M. O.; Mier, F.; Boeckler, F. M. Comparison of QM Methods for the Evaluation of Halogen−π Interactions for Large-Scale Data Generation. *J. Chem. Theory Comput.* **2025**, *21* (12), 6174−6183.

https://doi.org/10.1021/acs.jcim.5c02136
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

(47) Glick, Z. L.; Metcalf, D. P.; Glick, C. S.; Spronk, S. A.; Koutsoukas, A.; Cheney, D. L.; Sherrill, C. D. A physics-aware neural network for protein−ligand interactions with quantum chemical accuracy. *Chem. Sci.* **2024**, *15* (33), 13313−13324.
(48) Tong, S.; Lai, F. Artificial Neural Network-Based Approach for Surface Energy Prediction. In *Recent Advances in Neuromorphic Computing*; Bai, K. J.; Yi, Y., Eds.; IntechOpen, 2024.
(49) Unke, O. T.; Meuwly, M. PhysNet: A Neural Network for Predicting Energies, Forces, Dipole Moments, and Partial Charges. *J. Chem. Theory Comput.* **2019**, *15* (6), 3678−3693.
(50) Lu, C.; Li, S.; Lu, Z. Building energy prediction using artificial neural networks: A literature survey. *Energy Build.* **2022**, *262*, No. 111718.
(51) Bose, S.; Dhawan, D.; Nandi, S.; Sarkar, R. R.; Ghosh, D. Machine learning prediction of interaction energies in rigid water clusters. *Phys. Chem. Chem. Phys.* **2018**, *20* (35), 22987−22996.
(52) Umeno, Y.; Kubo, A. Prediction of electronic structure in atomistic model using artificial neural network. *Comput. Mater. Sci.* **2019**, *168*, 164−171.
(53) Zhang, Y.; Li, S.; Meng, K.; Sun, S. Machine Learning for Sequence and Structure-Based Protein−Ligand Interaction Prediction. *J. Chem. Inf. Model.* **2024**, *64* (5), 1456−1472.
(54) Shaw, R. A.; Hill, J. G. A Simple Model for Halogen Bond Interaction Energies. *Inorganics* **2019**, *7*, No. 19, DOI: 10.3390/inorganics7020019.
(55) Samuel, H. S.; Nweke-Maraizu, U.; Etim, E. E. Machine learning for characterizing halogen bonding interactions. *Fac. Nat. Appl. Sci. J. Sci. Innovations* **2023**, *5* (1), 102−114.
(56) Devore, D. P.; Shuford, K. L. Data and Molecular Fingerprint-Driven Machine Learning Approaches to Halogen Bonding. *J. Chem. Inf. Model.* **2024**, *64*, 8201−8214.
(57) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28* (1), 235−242.
(58) Afgan, E.; Allart, O.; The Galaxy, C.; et al. The Galaxy platform for accessible, reproducible, and collaborative data analyses: 2024 update. *Nucleic Acids Res.* **2024**, *52* (W1), W83−W94.
(59) Head-Gordon, M.; Pople, J. A.; Frisch, M. J. MP2 energy evaluation by direct methods. *Chem. Phys. Lett.* **1988**, *153* (6), 503−506.
(60) Møller, C.; Plesset, M. S. Note on an Approximation Treatment for Many-Electron Systems. *Phys. Rev.* **1934**, *46* (7), No. 618.
(61) *A Development of University of Karlsruhe and Forschungszentrum Karlsruhe GmbH*, TURBOMOLE V7.7.1; TURBOMOLE GmbH, 2007.
(62) Schäfer, A.; Huber, C.; Ahlrichs, R. Fully optimized contracted Gaussian basis sets of triple zeta valence quality for atoms Li to Kr. *J. Chem. Phys.* **1994**, *100* (8), 5829−5835.
(63) bwForCluster - JUSTUS 2, 2024. https://wiki.bwhpc.de/e/JUSTUS2.
(64) Feyereisen, M.; Fitzgerald, G.; Komornicki, A. Use of approximate integrals in ab initio theory. An application in MP2 energy calculations. *Chem. Phys. Lett.* **1993**, *208* (5), 359−363.
(65) Häser, M.; Ahlrichs, R. Improvements on the direct SCF method. *J. Comput. Chem.* **1989**, *10* (1), 104−111.
(66) Hättig, C. Geometry optimizations with the coupled-cluster model CC2 using the resolution-of-the-identity approximation. *J. Chem. Phys.* **2003**, *118* (17), 7751−7761.
(67) Hättig, C.; Hellweg, A.; Köhn, A. Distributed memory parallel implementation of energies and gradients for second-order Møller−Plesset perturbation theory with the resolution-of-the-identity approximation. *Phys. Chem. Chem. Phys.* **2006**, *8* (10), 1159−1169.
(68) Hättig, C.; Weigend, F. CC2 excitation energy calculations on large molecules using the resolution of the identity approximation. *J. Chem. Phys.* **2000**, *113* (13), 5154−5161.
(69) Hoffmann, R. An Extended Hückel Theory. I. Hydrocarbons. *J. Chem. Phys.* **1963**, *39* (6), 1397−1412.
(70) Weigend, F.; Ahlrichs, R. Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy. *Phys. Chem. Chem. Phys.* **2005**, *7* (18), 3297−3305.
(71) Weigend, F.; Häser, M. RI-MP2: first derivatives and global consistency. *Theor. Chem. Acc.* **1997**, *97* (1), 331−340.
(72) Weigend, F.; Häser, M.; Patzelt, H.; Ahlrichs, R. RI-MP2: optimized auxiliary basis sets and demonstration of efficiency. *Chem. Phys. Lett.* **1998**, *294* (1), 143−152.
(73) PyTorch, 2024. https://pytorch.org/.
(74) bwForCluster - BinAC, 2024. https://wiki.bwhpc.de/e/BinAC.
(75) *The PyMOL Molecular Graphics System*, Version 3.1.; Schrödinger, LLC, 2015.
(76) Mahalanobis, P. C. On the generalized distance in statistics *Proc. Natl. Inst. Sci. India* 1936; Vol. 2.

https://doi.org/10.1021/acs.jcim.5c02136
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Appendix D: Publication 4

Expanding and Enhancing Neural Network-Based QM-AI Models for the Accurate Prediction of Halogen-π Interaction Energies in Protein Contexts

Marc U. Engelhardt, Finn Mier, Markus O. Zimmermann, Frank M. Boeckler**;**

Journal of Chemical Information and Modeling, 2025

DOI: 10.1021/acs.jcim.5c03249

JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

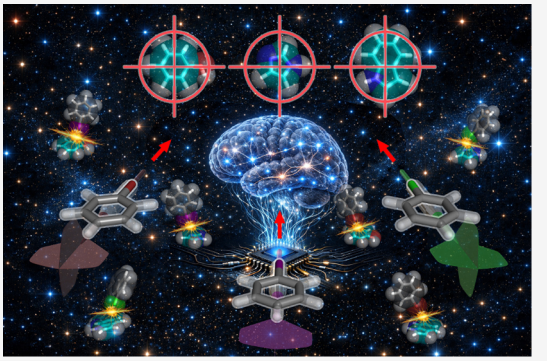
pubs.acs.org/jcim Article

Expanding and Enhancing Neural Network-Based QM-AI Models for the Accurate Prediction of Halogen-π Interaction Energies in Protein Contexts

Marc U. Engelhardt, Finn Mier, Markus O. Zimmermann, and Frank M. Boeckler*

Cite This: https://doi.org/10.1021/acs.jcim.5c03249 Read Online

ABSTRACT: In this study, we extend a previously introduced QM-AI strategy for predicting halogen···π interaction energies from a single aromatic model (representing phenylalanine) to multiple biologically relevant aromatic environments. Herein, neural network models were developed for halogen···π interactions involving phenol, imidazole, and indole, serving as model systems for the aromatic side chain residues of tyrosine, histidine, and tryptophan. Large, systematically generated datasets of halobenzene-aromatic system complexes (in total, over 18 million interaction geometries) were evaluated at the MP2/TZVPP level of theory and represented by compact geometric descriptors to train residue-specific neural networks. Across all systems, the models reproduce quantum-mechanical reference energies with high accuracy ($R^2 > 0.98$ and RMSE < 0.5 kJ/mol) within the targeted σ-hole interaction domain and retain robust performance on independent, randomly generated geometry and PDB-derived test sets. Model limitations are primarily associated with geometric arrangements outside the training distribution, such as π···π, C−H···π, or other non-σ-hole interaction motifs. By augmenting the training data with additional randomly generated geometries, model robustness and generalization were further improved without modifying the underlying network architecture. Overall, this work establishes a scalable and transferable QM-AI strategy for the rapid and accurate prediction of halogen···π interaction energies across diverse aromatic environments, enabling near-quantum-mechanical accuracy at negligible computational cost and supporting future applications in structure-based drug design.



INTRODUCTION

Understanding the individual contributions of different noncovalent interactions is crucial for elucidating biomolecular functions and guiding the rational design of potential therapeutics.[1−5] Halogen bonding (XB) is a directional noncovalent interaction between a halogen atom acting as an electrophilic site due to its σ-hole, a positive electrostatic region along the extension of the C-X bond (X = Cl, Br, or I), and a nucleophilic site, such as a lone pair or π-system.[6−10] The highly anisotropic electron distribution around the halogen atoms results in a pronounced lateral electron density oriented perpendicular to the R-X bond axis. When the substituent R exerts a pronounced electron-withdrawing influence on the halogen (X), such tuning effects lead to a substantial increase in halogen-bond strength.[11−16] In biomolecular systems, halogen bonds have been increasingly recognized as relevant contributors to protein−ligand binding and drug-target recognition.[17−27] Aromatic π-systems, heteroaromatic residues, and backbone carbonyl groups can act as halogen bond acceptors, giving rise to well-defined interaction motifs in protein−ligand complexes. Consequently, halogen substitution has become an established strategy in medicinal chemistry to modulate binding affinity, selectivity, and pharmacokinetic properties.[28−35]

Computational chemistry has played a central role in elucidating both the structural preferences and energetic characteristics of halogen bond interactions.[36−38] Previous studies have focused predominantly on halogen bonds formed with classical nucleophilic sites commonly found in protein-binding pockets, such as backbone carbonyl oxygen atoms,[39,40] peptide π-surfaces,[41] heteroatoms in amino acid side chains,[42,43] carboxylate and carboxamide functionalities,[44] and coordinated or structural water molecules.[45,46]

Recently, we investigated halogen bonding to the electron-rich π-system of the aromatic amino acid side chain of phenylalanine using high-level quantum-mechanical calculations. We demonstrated that MP2 calculations with sufficiently large basis sets (TZVPP) reproduce CCSD(T)/CBS reference

Received: December 31, 2025
Revised: March 13, 2026
Accepted: April 6, 2026

https://doi.org/10.1021/acs.jcim.5c03249
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

interaction energies for halogen···π complexes with high accuracy, while reducing computational cost by approximately 2 orders of magnitude relative to CCSD(T).[47] This efficiency gain enabled the generation of large, high-quality reference datasets suitable for systematic analysis. Building on this foundation, we subsequently trained neural network models on the MP2-derived data to predict halogen···π interaction energies directly from geometric descriptors.[48] These models achieved excellent predictive accuracy at negligible computational cost, corresponding to an overall acceleration of roughly 10 orders of magnitude compared to CCSD(T) calculations. The results demonstrated that directional σ-hole interactions can be reliably captured using well-chosen geometric features by exploiting a methodological and hierarchical "double-jump" strategy from CCSD(T) to MP2 and ultimately to neural networks, providing a practical alternative to explicit quantum-mechanical evaluations of halogen···π interactions.

In the present study, we build directly on this framework and extend it beyond a single aromatic reference system, implementing aromatic environments representative of the remaining amino acid side chains. Specifically, we investigate halogen···π interactions involving phenol, imidazole, and indole, serving as model systems for the aromatic side chain residues of tyrosine, histidine, and tryptophan, respectively. For histidine, only the neutral imidazole form was considered, while the protonated imidazolinium state was excluded. Although histidine can be protonated in certain protein environments (e.g., kinase active sites), explicitly accounting for variable protonation states would substantially increase the model complexity and require extensive site-specific preprocessing. Moreover, protonated histidine (imidazolinium) and related cationic π-systems, such as the guanidinium group of arginine, represent a distinct class of positively charged aromatic interaction motifs that are beyond the scope of the present work. From a physicochemical perspective, positively charged histidine is also expected to be less favorable for σ-hole-driven halogen bonding, as electrostatic attraction would preferentially promote charge-assisted hydrogen bonding or interactions with the lateral electron-rich belt of the halogen rather than directional σ-hole contacts, potentially leading to reoriented binding geometries. In contrast, neutral imidazole provides a more plausible acceptor environment for stabilizing genuine halogen···π interactions. A systematic treatment of positively charged π-systems, including both imidazolinium and guanidinium motifs, therefore represents a related and promising direction for future investigation. For each system, large and systematically generated datasets of approximately 6 million interaction geometries between halobenzenes (chlorobenzene, bromobenzene, and iodobenzene) and the corresponding aromatic acceptor were constructed and evaluated at the MP2/TZVPP level of theory. As in our previous work, the interaction geometries were sampled on structured grids defined by halogen···π-plane distances (from $d_{\min(X\cdots\pi\text{-plane})}$ = 2.75 Å to $d_{\max(X\cdots\pi\text{-plane})}$ = 4.50 Å) and angular constraints (maximum deviation of 40° between the C-X bond vector and the π-plane normal) chosen to selectively represent directional σ-hole interactions while minimizing contributions from competing motifs such as $\pi\cdots\pi$ or C−H···π contacts. The resulting quantum-mechanical reference data were used to train dedicated neural network models for each aromatic system in a supervised learning approach. Model performance was assessed using both test sets of randomly generated geometries and geometries derived from protein crystal structures of the Protein Data Bank (PDB),[49] enabling a direct evaluation of transferability to biologically relevant examples. Here again, only neutral histidine/imidazole motifs were retained, as reliable inclusion of protonated histidine would require prior protonation-state assignment, binding-site environment analysis, and optimization of hydrogen-bond networks for each structure, which is computationally prohibitive on the scale considered in this study.

In addition to extending the framework to multiple aromatic side-chain models, the present study explicitly examines the modular expandability of the trained neural network models. A central premise of the approach is that model robustness and transferability can be systematically improved by incorporating additional training data without modifying the underlying network architecture. To assess this capability, we further generated approximately 100,000 randomly sampled interaction geometries for each aromatic system and halogen type and employed them in a retraining step. These augmented datasets were designed to broaden the sampled feature space beyond the structured grids used for the initial training, thereby enabling a targeted evaluation of how expanded training distributions influence predictive performance, generalization to previously challenging geometries, and transferability to PDB-derived interaction motifs. By combining systematic data generation, controlled retraining, and evaluation on both randomly sampled and biologically derived test sets, this study aims to establish a scalable and adaptable machine-learning framework for the efficient prediction of halogen···π interaction energies across diverse aromatic environments. It should be noted that the present models predict gas-phase interaction energies (ΔE) rather than binding free energies (ΔG). In practical structure-based drug design workflows, solvation, desolvation penalties, and entropic contributions are accounted for by the surrounding docking or scoring framework, while the present models provide an interaction-specific description of directional halogen···π interactions. This represents a further step toward integration into structure-based drug design workflows, including molecular docking and scoring applications.

RESULTS AND DISCUSSION

Model Training and Validation

First, we conducted single-point calculations of the generated interaction geometries. Energies were calculated by using the supermolecular approach (eq 1). Features were derived from the individual interaction geometries. A complete list of all features of the individual model systems, along with descriptions and a more detailed schematic of the feature definitions, is provided in the Supporting Information (Figure S1, Tables S1 and S2). To ensure the feature space is translationally and rotationally independent of the underlying coordinate system, we chose pairwise distances and angles. For the training process, the resulting datasets were partitioned into training, validation, and test subsets (for more details, see Materials and Methods). Hyperparameter search and training processes were performed in a stratified 5-fold cross-validation with an 80%/20% test/validation split. Final model training on the best-performing model configuration was conducted using a slightly adapted training subset (95% of the data), with evaluation on the corresponding validation subset (5% of the data) at the end of each epoch. Training was terminated upon satisfaction of the early stopping criterion. Model evaluation of

https://doi.org/10.1021/acs.jcim.5c03249
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

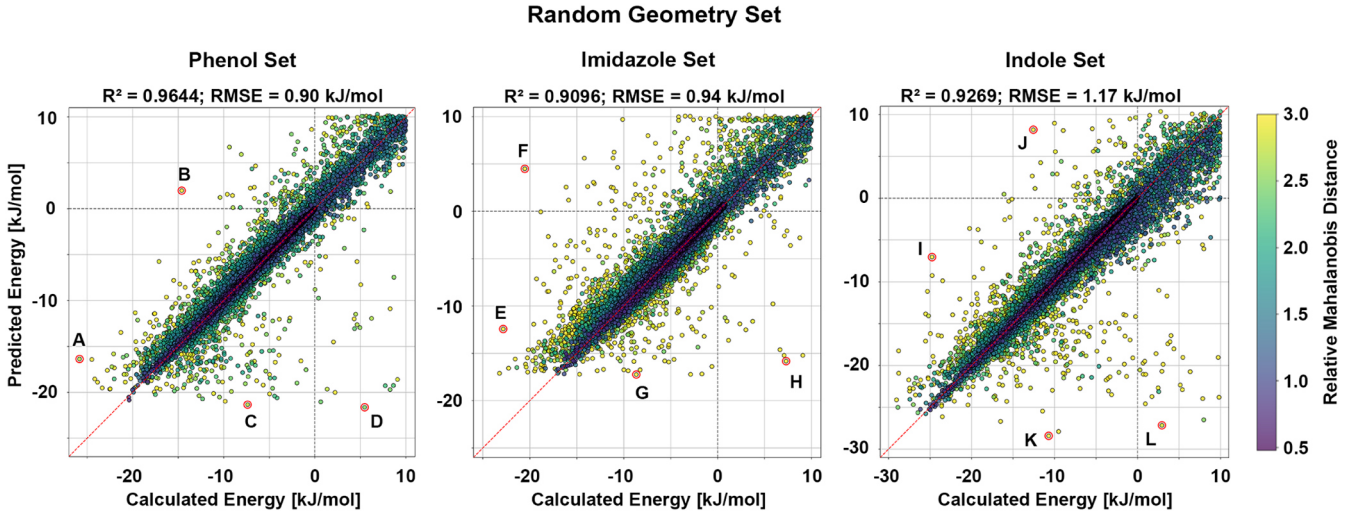


Figure 1. Model performance on the random geometry set. Calculated adduct formation energies are plotted against the corresponding predicted values for the phenol, imidazole, and indole models. The phenol model achieves an R^2 = 0.9644 with an RMSE = 0.90 kJ/mol, the imidazole model achieves an R^2 = 0.9096 with an RMSE = 0.94 kJ/mol, and the indole model achieves an R^2 = 0.9269 with an RMSE = 1.17 kJ/mol. The red dashed line denotes perfect agreement between the calculated and predicted energies, while the gray dashed lines mark the transition between negative and positive energies. Data points are colored according to their relative MD, as indicated by the color scale. Data points highlighted by red circles and labeled A−L correspond to selected outliers and are discussed in detail in Figure 2.

the withheld test sets shows excellent accuracy across all three models, with only very few large energy deviations ($\Delta\Delta E$, eq 2). For each system, the vast majority of predictions fall within ±0.5 kJ/mol, and only a small fraction exceeds this range, with 1.49% for phenol, 5.04% for imidazole, and 2.02% for indole. For phenol, the maximum and minimum deviations were +4.08 and −3.11 kJ/mol, respectively, yielding an R_{phenol}^2 = 0.9983 and an $RMSE_{phenol\text{-}test}$ = 0.15 kJ/mol, demonstrating that predictive performance remains strong even on unseen data (Figure S2a/b).

Similarly, the imidazole model exhibited maximum and minimum deviations of +8.92 and −7.84 kJ/mol, respectively, achieving an $R_{imidazole}^2$ = 0.9919 and an $RMSE_{imidazole\text{-}test}$ = 0.28 kJ/mol (Figure S2c/d). For indole, the deviations ranged from +3.76 to −2.95 kJ/mol, with an R_{indole}^2 = 0.9982 and an $RMSE_{indole\text{-}test}$ = 0.17 kJ/mol (Figure S2e/f). Despite these promising results, the performance may be somewhat inflated due to similarities between the training and test sets, potentially introducing data leakage, as both likely occupy the same region of feature space. Therefore, a more rigorous assessment requires evaluating the models on entirely novel data that fall outside the distribution of the training set to truly estimate their generalization capabilities. In addition to such fully unseen, out-of-distribution samples, it is equally important to assess data points that lie within the training distribution but involve interpolated parameter combinations. This ensures that the models are not merely memorizing known values but can also generalize smoothly across the interior of the feature space. Therefore, a two-step evaluation scheme of the individual models' test datasets was applied. In the first step (i), a subset containing only data points with features lying within the training distribution (2.75 Å ≤ $d_{X\cdots\pi\text{-plane}}$ ≤ 4.5 Å; $\alpha_{C\text{-}X\cdots\perp(\pi\text{-plane})}$ ≤ 40°) was analyzed. In the second step (ii), distance and angle constraints were dropped, and the full test dataset was evaluated. This removal of the restrictions was intended to test and identify the models' limitations and ability to generalize on unseen data on the edges or outside of the feature space covered by the underlying training distribution.

Predicting Adduct Formation Energies of Unseen Datasets

To rigorously assess the generalization capability of the trained models, we generated independent test sets comprising interaction geometries with randomly sampled translational and rotational parameters for each molecular system. Each test set contained 60,000 structures (20,000 per halogen). Single-point interaction energies were computed at the MP2/TZVPP level of theory, and repulsive geometries with interaction energies exceeding +10 kJ/mol were discarded. The resulting curated datasets contained N_{phenol} = 56,591, $N_{imidazole}$ = 57,297, and N_{indol} = 55,415 complexes. Geometric features were extracted from each complex. A detailed table incorporating the individual data points, energies E_{calc} and E_{pred}, as well as the difference between both, is provided in spreadsheet format in the Supporting Information. To examine the dependence of model performance on the amount of training data, we constructed learning curves by training the models on progressively larger fractions of the original training set. Model accuracy was evaluated both on the internal test split and on the independent random-geometry test set described above. A detailed representation of the learning curves is provided in the Supporting Information (Figure S7). In both evaluations, the prediction errors decrease rapidly with increasing training set size and approach a plateau once approximately 30−40% of the data is included. Beyond this point, further enlargement of the training set yields only marginal improvements in MAE and RMSE. These results indicate that the chosen feature representation efficiently captures the relevant geometry−energy relationships, while the full dataset primarily contributes to improved robustness and coverage of the interaction space.

Phenol Model Evaluation

The halobenzene···phenol model was evaluated using the two-step protocol described above. In the constrained subset (step i, N = 16,062), which contains only geometries lying within the feature space sampled during training (2.75 Å ≤ $d_{X\cdots\pi\text{-plane}}$ ≤ 4.5 Å, $\alpha_{C\text{-}X\cdots\perp(\pi\text{-plane})}$ ≤ 40°), the model attained excellent

https://doi.org/10.1021/acs.jcim.5c03249
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

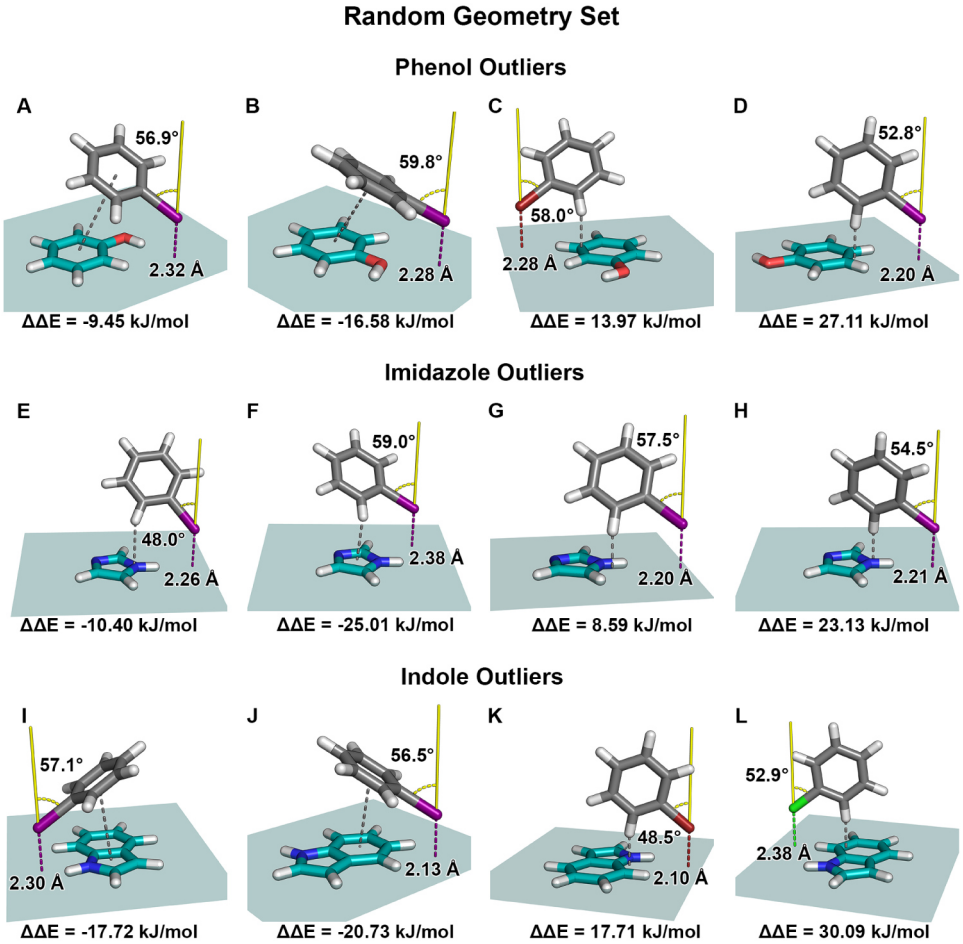


Figure 2. Interaction geometries for selected outliers from the random geometry set involving phenol, imidazole, and indole, as identified in the scatter plot in Figure 1. These structures correspond to data points exhibiting large deviations between calculated and predicted adduct formation energies. Geometries A−D depict the relative orientation of the halobenzene (gray) with respect to the phenol π-system (teal). Shown are the halogen distance $d_{X\cdots\pi\text{-plane}}$ (dashed line, colored according to the halogen) to the π-system plane in Å, the torsion angle $\alpha_{C\text{-}X\cdots\perp(\pi\text{-plane})}$ (yellow) between the C-X bond vector and the normal to the π-system plane in degrees, and the corresponding energy difference $\Delta\Delta E$. The gray dashed line indicates the type of interaction that may contribute to the observed prediction error. For clarity, the teal plane also illustrates the extent of the training grid used in the corresponding model. Geometries E−H show halobenzene interactions with imidazole, while geometries I−L show interactions with indole. Outliers of the phenol dataset: (A) Iodobenzene interaction with $d_{I\cdots\pi\text{-plane}}$ = 2.32 Å and $\alpha_{C-I\cdots\perp(\pi\text{-plane})}$ = 56.9°. The gray dashed line indicates a $\pi\cdots\pi$-interaction. (B) Iodobenzene interaction with $d_{I\cdots\pi\text{-plane}}$ = 2.28 Å and $\alpha_{C-I\cdots\perp(\pi\text{-plane})}$ = 59.8°. The gray dashed line indicates a $\pi\cdots\pi$-interaction. (C) Bromobenzene interaction with $d_{Br\cdots\pi\text{-plane}}$ = 2.28 Å and $\alpha_{C-Br\cdots\perp(\pi\text{-plane})}$ = 58.0°. The gray dashed line indicates a C−H$\cdots\pi$ contact. (D) Iodobenzene interaction with $d_{I\cdots\pi\text{-plane}}$ = 2.20 Å and $\alpha_{C-I\cdots\perp(\pi\text{-plane})}$ = 52.8°. The gray dashed line indicates a C−H$\cdots\pi$ contact. Outliers of the imidazole dataset: (E) Iodobenzene interaction with $d_{I\cdots\pi\text{-plane}}$ = 2.26 Å and $\alpha_{C-I\cdots\perp(\pi\text{-plane})}$ = 48.0°. The gray dashed line indicates a C−H$\cdots\pi$ contact. (F) Iodobenzene interaction with $d_{I\cdots\pi\text{-plane}}$ = 2.38 Å and $\alpha_{C-I\cdots\perp(\pi\text{-plane})}$ = 59.0°. The gray dashed line indicates a C−H$\cdots\pi$ contact. (G) Iodobenzene interaction with $d_{I\cdots\pi\text{-plane}}$ = 2.20 Å and $\alpha_{C-I\cdots\perp(\pi\text{-plane})}$ = 57.5°. The gray dashed line indicates a C−H$\cdots\pi$ contact. (H) Iodobenzene interaction with $d_{I\cdots\pi\text{-plane}}$ = 2.21 Å and $\alpha_{C-I\cdots\perp(\pi\text{-plane})}$ = 54.5°. The gray dashed line indicates a C−H$\cdots\pi$ contact. Outliers of the indole dataset: (I) Iodobenzene interaction with $d_{I\cdots\pi\text{-plane}}$ = 2.30 Å and $\alpha_{C-I\cdots\perp(\pi\text{-plane})}$ = 57.1°. The gray dashed line indicates a $\pi\cdots\pi$ interaction. (J) Iodobenzene interaction with $d_{I\cdots\pi\text{-plane}}$ = 2.13 Å and $\alpha_{C-I\cdots\perp(\pi\text{-plane})}$ = 56.5°. The gray dashed line indicates a $\pi\cdots\pi$ interaction. (K) Bromobenzene interaction with $d_{Br\cdots\pi\text{-plane}}$ = 2.10 Å and $\alpha_{C-Br\cdots\perp(\pi\text{-plane})}$ = 48.5°. The gray dashed line indicates a C−H$\cdots\pi$ contact. (L) Chlorobenzene interaction with $d_{Cl\cdots\pi\text{-plane}}$ = 2.38 Å and $\alpha_{C-Cl\cdots\perp(\pi\text{-plane})}$ = 52.9°. The gray dashed line indicates a C−H$\cdots\pi$ interaction.

accuracy (R^2 = 0.9935, RMSE = 0.27 kJ/mol), essentially matching the performance observed during training with low energy deviations between $\Delta\Delta E$ = +3.21 and −3.65 kJ/mol, and 5.51% lying beyond ±0.5 kJ/mol (Figure S3).

Evaluating the model on the full dataset (step ii, N = 56,591), where all geometric constraints were removed, resulted in a modest decrease in accuracy but still strong predictive ability (R^2 = 0.9644, RMSE = 0.90 kJ/mol). The observed deviations ranged from $\Delta\Delta E$ = +27.98 to −16.58 kJ/mol, with 12.41% lying beyond ±0.5 kJ/mol. To better understand the origins of these deviations, a parity plot of E_{calc} and E_{pred} was generated and colored by Mahalanobis distance[50] (MD) (Figure 1). The MD reflects how far each geometry lies from the center of the training feature distribution, accounting for feature correlations through the covariance matrix (eq 3). Most points lie close to the diagonal and exhibit low MDs, indicating close agreement between predicted and reference energies for structures similar to those encountered during training. In contrast, points with larger errors typically show elevated MDs, demonstrating a clear link between reduced predictive performance and increasing dissimilarity from the training distribution. Four representative outliers (A−D) were selected for closer inspection (Figure 2). It should be noted that these subsequently discussed examples are some of the

https://doi.org/10.1021/acs.jcim.5c03249
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Table 1. Overview of Energy Values and Geometric Parameters for Outlier Structures of the Random Geometry Set of Phenol (A–D), Imidazole (E–H), and Indole (I–L), Highlighted in Figure 1 and Depicted in Figure 2[a]

	halogen	π-system	ΔE_{calc} [kJ/mol]	ΔE_{pred} [kJ/mol]	ΔΔE [kJ/mol]	$d_{X\cdots\pi\text{-plane}}$ [Å]	$\alpha_{(C\text{-}X\cdots\perp(\pi\text{-plane}))}$ [°, deg]
A	I	phenol	−25.83	−16.38	−9.45	2.32	56.88
B	I	phenol	−14.61	1.97	−16.58	2.28	59.85
C	Br	phenol	−7.38	−21.35	13.97	2.28	57.96
D	I	phenol	5.47	−21.63	27.11	2.20	52.76
E	I	imidazole	−22.83	−12.43	−10.40	2.26	48.00
F	I	imidazole	−20.53	4.49	−25.01	2.38	58.97
G	I	imidazole	−8.65	−17.24	8.59	2.20	57.51
H	I	imidazole	7.30	−15.83	23.13	2.21	54.55
I	I	indole	−24.76	−7.04	−17.72	2.30	57.57
J	I	indole	−12.55	8.18	−20.73	2.13	56.54
K	Br	indole	−10.70	−28.41	17.71	2.10	48.47
L	Cl	indole	2.94	−27.15	30.09	2.38	52.94

[a]The halogen symbolizes the interacting halobenzene. Values of calculated and predicted energies, as well as the difference between both, are given in kJ/mol. Distance values $d_{X\cdots\pi\text{-plane}}$ between the halogen and the π-plane are given in Å. Angle values between the C-X vector and the normal to the π-plane are given in degrees [°, deg].

very few strong outliers, highlighting "worst-case" scenarios. Calculated and predicted adduct formation energies, energy differences, distances $d_{X\cdots\pi\text{-plane}}$ between the halogen and the π-plane, as well as the angle between the C-X vector and the normal to the π-plane $\alpha_{C\text{-}X\cdots\perp(\pi\text{-plane})}$ of these phenol outliers are listed in Table 1. Outliers can arise for two principal reasons: (a) the geometry resides outside the feature space spanned by the training data, requiring the model to extrapolate; or (b) the geometry lies within the training distribution but is still poorly predicted, implying limitations in the model architecture or training strategy.

Example A corresponds to a complex formed between iodobenzene and phenol and exhibits a highly compact interaction geometry. The iodine atom is positioned at a distance of $d_{I\cdots\pi\text{-plane}}$ = 2.32 Å and an angle of $\alpha_{C-I\cdots\perp(\pi\text{-plane})}$ = 56.9°. For this configuration, the predicted adduct formation energy deviates strongly from the reference value, yielding an energy difference of $\Delta\Delta E$(A) = −9.45 kJ/mol (ΔE_{calc} = −25.83 kJ/mol, ΔE_{pred} = −16.38 kJ/mol), indicating a pronounced underestimation of the interaction strength by the model. This discrepancy can be attributed to the fact that the corresponding geometric descriptors lie well outside the region of the feature space sampled during training. This is reflected by a Mahalanobis distance that is more than three times higher than that of the distribution center. Comparison with the training set confirms this assumption because the iodine···π-plane separation in Example A is markedly shorter than the minimum distance encountered during training ($d_{min(X\cdots\pi\text{-plane})}$(training) = 2.75 Å), while the associated torsion angle substantially exceeds the largest value present in the training data ($\alpha_{max(C\text{-}X\cdots\perp(\pi\text{-plane}))}$(training) = 40°). In addition, the proximity of the halogen atom to the phenol ring, and in particular to the hydroxyl substituent, may introduce stabilizing contributions not adequately captured by the model, potentially arising from interaction between the negatively charged region surrounding the halogen and the nearby hydrogen atom. The mutual arrangement of the aromatic rings may further stabilize the complex through a π···π interaction.

Example B exhibits a comparable trend, with the predicted interaction energy again markedly underestimated ($\Delta\Delta E$(B) = −16.58 kJ/mol). This discrepancy arises from a qualitative misclassification of the interaction, as the calculated and predicted energies differ not only in magnitude but also in sign (ΔE_{calc} = −14.61 kJ/mol, ΔE_{pred} = 1.97 kJ/mol), indicating that the stabilizing π···π contribution is not adequately represented by the model. As in Example A, the associated distance and angle values fall well outside the range sampled during training, indicating a closely related but distinct geometric representation. In this configuration, the mutual alignment of the aromatic rings enhances the π···π interaction character even further. The resulting prediction error highlights a broader limitation of the model, which appears insufficiently able to generalize such highly compact, strongly tilted geometries that enable favorable π···π or X···H interaction motifs, thereby leading to a systematic underestimation of their stabilizing energetic contributions.

Examples C and D show the opposite trend compared to A and B, with the interaction energies strongly overestimated by the model. In Example C, the calculated energy is moderately stabilizing (ΔE_{calc} = −7.38 kJ/mol), whereas the model predicts a substantially stronger interaction (ΔE_{pred} = −21.35 kJ/mol), resulting in a large positive deviation ($\Delta\Delta E$(C) = 13.97 kJ/mol). An even more pronounced discrepancy is observed for Example D, where the calculated energy indicates a slightly repulsive interaction (ΔE_{calc} = 5.47 kJ/mol), while the model predicts a strongly attractive complex (ΔE_{pred} = −21.63 kJ/mol).

In both geometries, the halogen approaches the phenol π-system at very short distances and with pronounced tilt angles, placing these structures outside the range of configurations represented in the training data. While such compact arrangements may still resemble stabilizing contacts, the reference calculations indicate that short-range repulsive contributions can become significant in these cases. The inability of the model to capture this repulsive part suggests that it primarily extrapolates attractive interactions from short contact distances, thereby failing to account for steric and electronic repulsion at close approach.

In the present case, all four examples fall into category (a), exhibiting features not covered during training and correspondingly high MD values. Because these atypical and often repulsive interaction geometries were not included in the training set, the model lacks the necessary information to penalize them appropriately. The same applies to the recognition of π···π interactions and their beneficial contribution. While accurately modeling these configurations is not

https://doi.org/10.1021/acs.jcim.5c03249
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

essential, since they are typically identified and filtered out earlier or handled by dedicated repulsion terms, future models may benefit from including representative repulsive σ-hole or $\pi\cdots\pi$ arrangements to further enhance robustness and transferability.

Imidazole Model Evaluation

The imidazole model was evaluated using the same two-step procedure. For the constrained subset (step (i), N = 16,075), the model achieved high accuracy (R^2 = 0.9848, RMSE = 0.34 kJ/mol) with $\Delta\Delta E$ between +3.71 kJ/mol and −3.17 kJ/mol (9.42% beyond ±0.5 kJ/mol). When applied to the full test set (step (ii), N = 57,297), the performance remained strong with R^2 = 0.9096 and RMSE = 0.94 kJ/mol, although the spread of deviations widened, with $\Delta\Delta E$ values ranging from +25.01 kJ/mol to −24.29 kJ/mol (19.28% beyond ±0.5 kJ/mol).

A corresponding parity plot of the imidazole results is shown in Figure 1. Similar to the phenol model, most data points cluster along the parity line, while a small number of higher-MD structures exhibit larger prediction errors. Four extreme outliers (labeled E−H) are examined in detail (Figure 2). Calculated and predicted adduct formation energies, energy differences, distances $d_{X\cdots\pi\text{-plane}}$ between the halogen and the π-plane, as well as the angle between the C-X vector and the normal to the π-plane $\alpha_{C\text{-}X\cdots\perp(\pi\text{-plane})}$ of these imidazole outliers, are listed in Table 1.

Examples E−H exhibit closely related interaction geometries characterized by strongly tilted halobenzene orientations and halogen positions located near the boundaries of the model's training grid. In all four cases, the halogen atom resides outside the π-plane, and the dominant stabilizing motif is limited to C−H$\cdots\pi$ contacts rather than direct halogen$\cdots\pi$ interactions. The primary geometric differences among these structures arise from the specific positioning of the halogen, which, in turn, governs the relative placement of the C−H$\cdots\pi$ and H$\cdots$X contacts. Despite their overall similarity, three of the four geometries (E−G) correspond to attractive interactions, whereas example H is predicted to be repulsive by the reference calculation. A particularly instructive comparison can be made between examples E and H, which share nearly identical orientations but differ subtly in distance and angular values. In example E, the slightly larger contact distance and smaller interaction angle give rise to a strongly stabilizing interaction, while the modest geometric changes observed in example H shift the balance toward repulsion with a shorter distance and larger angle. This sharp transition from attractive to repulsive behavior occurs over a narrow region of geometric space that is sparsely sampled in the training data and is therefore not reliably captured by the model. It should be emphasized that the geometries discussed here represent only a limited subset, selected from the "worst-case" identified outliers, and are not intended to be exhaustive. They serve as illustrative examples highlighting characteristic failure sources of the model rather than a comprehensive description of all deviations present in the test set. In particular, although not shown explicitly, the imidazole-derived data also contain outlier geometries involving alternative interaction motifs, such as $\pi\cdots\pi$ contacts, which may give rise to similar prediction challenges. In summary, the imidazole model demonstrates performance analogous to the phenol model, with closely related error characteristics and limitations, but retains good overall agreement between calculated and predicted energies.

Indole Model Evaluation

The indole model showed similarly strong performance under the two-step evaluation framework. Within the constrained region (step (i), N = 15,211), the model achieved excellent agreement with reference data (R^2 = 0.9953, RMSE = 0.28 kJ/mol) and notably low energy deviations between +2.72 kJ/mol and −2.36 kJ/mol (6.94% beyond ±0.5 kJ/mol). Evaluation on the unconstrained full dataset (step (ii), N = 55,415) resulted in a decrease to R^2 = 0.9269 and RMSE = 1.17 kJ/mol, with deviations of $\Delta\Delta E$ between +34.47 kJ/mol and −20.73 kJ/mol (23.36% beyond ±0.5 kJ/mol). This indicates that the indole feature space contains a broader distribution of challenging geometries when the distance and angle constraints are removed and could be due to the larger size of the acceptor system.

The corresponding parity plot of the indole results (Figure 1) again highlights four outliers (I−L), which exhibit the largest discrepancies. All individual values of these four indole outliers can be found in Table 1. As with the phenol and imidazole models, these structures correspond to geometries lying well outside the training distribution, consistent with elevated MD values.

Examples I−L illustrate characteristic limitations of the indole model that are closely related to those observed for phenol and imidazole. Examples I and J feature pronounced $\pi\cdots\pi$ interaction motifs and correspond to strongly stabilizing reference energies that are not adequately captured by the model. Although both geometries exhibit similarly large tilt angles, small variations in the intermolecular distance lead to qualitatively different predictions. While example I, with a slightly larger separation, is still predicted to be attractive, example J, characterized by a shorter contact distance, is incorrectly classified as repulsive. This behavior indicates limited generalizability of the model with respect to distance variations in compact $\pi\cdots\pi$ arrangements and distances beyond the feature space during training.

In contrast, examples K and L are dominated by C−H$\cdots\pi$ interactions at short contact distances and high angular values. In these cases, the balance between stabilizing H$\cdots$X interactions and short-range repulsive contributions is particularly delicate. A narrow geometric transition separates attractive configurations, where favorable H$\cdots$X contacts prevail, from repulsive arrangements dominated by steric C−H clashes. This subtle crossover occurs over a region of feature space that is insufficiently represented in the training data and, is therefore, not reliably reproduced by the model.

Thus, the general trend across all three models is that prediction quality decreases as the geometries diverge further from the region sampled during training, particularly in repulsive or otherwise atypical configurations such as $\pi\cdots\pi$ or C−H$\cdots\pi$ arrangements.

PDB Scan for Halogen$\cdots\pi$ Interactions in Crystal Structures

In a previous study, we performed a comprehensive PDB scan to identify halogen$\cdots\pi$ interactions involving phenylalanine.[48] Building on that work, we expanded the analysis to focus on the aromatic side chains of tyrosine, histidine, and tryptophan. 239,149 crystal structures (as of July 2025) were analyzed, with 9810 unique PDB IDs containing chlorine-, bromine-, or iodine-bearing ligands. Phenylalanine accounts for over 40% of all observed contacts (already analyzed in our earlier work),[48] followed by tyrosine at 29.36%. Histidine (15%) and tryptophan (12%) are less frequently encountered but still

https://doi.org/10.1021/acs.jcim.5c03249
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

represent relevant biological environments for σ-hole interactions. This result is not unexpected, as Trp is encoded only by one DNA triplet in comparison to two DNA triplets for Phe, Tyr, and His. Recent analyses of amino acid frequencies in human PDB structures and human proteins in the UniProt database highlight that Trp is about half as frequent as His, while Phe and Tyr are slightly more frequent than His.[51] While the trends are quite similar in our results, it is evident that His is less frequently found in close proximity to halogens compared to its statistical occurrence in human proteins of the pdb. However, this is most likely due to the multitude of other interactions His can exhibit, such as hydrogen bonding, charged interactions, or metal coordination. Results of the scan and a more detailed analysis of the individual amino acid values are shown in Table 2.

Table 2. PDB Scan Results for Halogen···π Contacts and Halogen···π Interactions[a]

Addressed AA	Halogen	Halogen···π contacts within 5 A (% of total)	Halogen···π interactions after applied filters (% of contacts)
Histidine	Cl, Br, I	3597 (15.28%)	661 (18.38%)
Phenylalanine	Cl, Br, I	10174 (43.23%)	1114 (10.95%)
Tryptophan	Cl, Br, I	2855 (12.13%)	384 (13.45%)
Tyrosine	Cl, Br, I	6910 (29.36%)	1152 (16.67%)
Total		23536	3311 (14.07%)
Tyrosine	Cl	5275 (76.34%)	856 (74.30%)
Tyrosine	Br	1232 (17.83%)	225 (19.54%)
Tyrosine	I	403 (5.83%)	71 (6.16%)
Histidine	Cl	2966 (82.46%)	602 (91.07%)
Histidine	Br	489 (13.59%)	50 (7.56%)
Histidine	I	142 (3.95%)	9 (1.37%)
Tryptophan	Cl	2150 (75.31%)	296 (77.08%)
Tryptophan	Br	567 (19.86%)	76 (19.79%)
Tryptophan	I	138 (4.83%)	12 (3.13%)

[a]Results of Interactions with Tyrosine, Histidine, and Tryptophan are Reported for Each Halogen Separately.

To isolate interactions characteristic of the halogen's σ-hole engagement, we applied distance and angular criteria and additionally required the C-X bond vector of the ligand to point toward the π-plane, an orientation consistent with the formation of a directional σ-hole contact. After applying these geometric filters, we identified N_{TYR} = 1152, N_{HIS} = 661, and N_{TRP} = 384 halogen···π interactions for further analysis.

For consistent comparison across systems, a matched-molecular-pair approach was employed. Each ligand halogen donor was substituted with the corresponding halobenzene model system. The halogen atom and C-X vector were aligned exactly with the original ligand geometry, and the benzene ring was superimposed onto the ligand's aromatic ring to preserve orientation. Likewise, the amino acid side chain was truncated at the C_γ−C_β bond and replaced with the protonated, optimized model system by aromatic-ring superposition. This substitution strategy necessarily neglects ligand-specific electronic tuning of halogen strength but ensures consistent geometric comparison. MP2/TZVPP single-point interaction energies were computed using TURBOMOLE. Repulsive geometries with interaction energies exceeding +10 kJ/mol were discarded, resulting in final datasets of N_{TYR} = 1142, N_{HIS} = 660, and N_{TRP} = 384 geometries. Although the datasets describe halobenzene-phenol, -imidazole, and -indole interactions, we refer to them as the "tyrosine," "histidine," and "tryptophan" datasets, respectively, to distinguish the PDB-derived sets from the randomly generated ones.

Model Evaluation on PDB-Derived Sets

Addressing the Side Chain of Tyrosine. For the 1142 PDB-derived halogen···TYR geometries (denoted as "tyrosine set"), features were extracted and submitted to the phenol-based NN model. In the restricted evaluation (step (i) 578 structures lie within the training distribution (2.75 Å ≤ $d_{X\cdots\pi\text{-plane}}$ ≤ 4.5 Å, $\alpha_{C\text{-}X\cdots\perp(\pi\text{-plane})}$ ≤ 40°). Within this domain, the model achieved excellent accuracy ($R^2_{TYR(i)}$ = 0.9904, $RMSE_{TYR(i)}$ = 0.30 kJ/mol), with deviations confined to +0.71 to −1.58 kJ/mol, with 8.13% beyond ±0.5 kJ/mol (Figure S4).

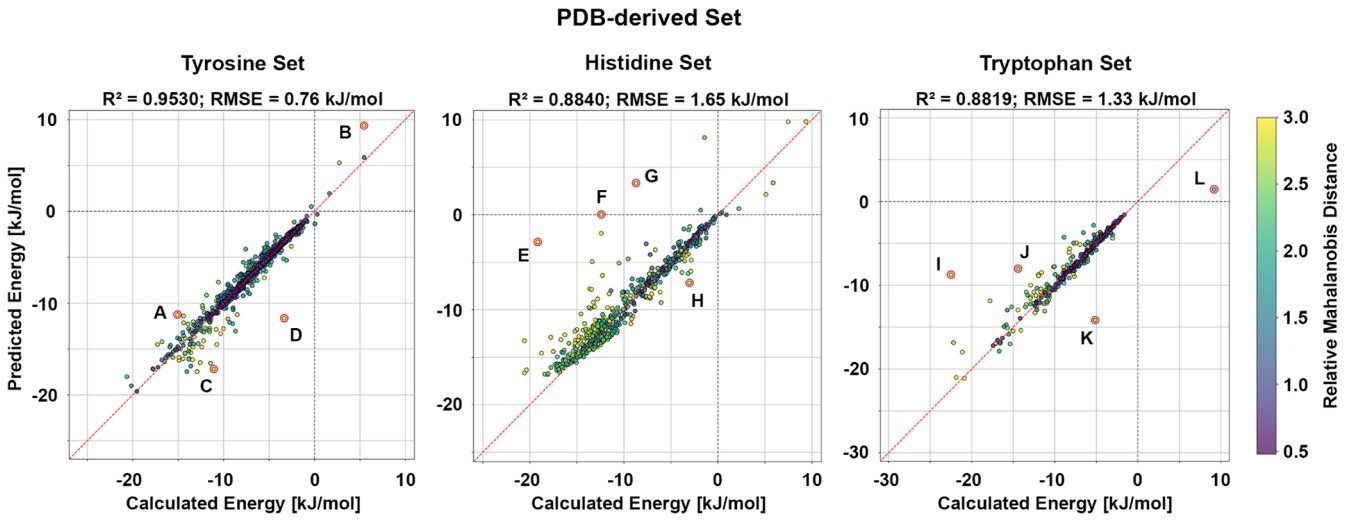


Figure 3. Model performance on the PDB-derived set. Calculated adduct formation energies are plotted against the corresponding predicted values for the phenol, imidazole, and indole models. The phenol model achieves an R^2_{TYR} = 0.9530 with an $RMSE_{TYR}$ = 0.76 kJ/mol, the imidazole model achieves an R^2_{HIS} = 0.8840 with an $RMSE_{HIS}$ = 1.65 kJ/mol, and the indole model achieves an R^2_{TRP} = 0.8819 with an $RMSE_{TRP}$ = 1.33 kJ/mol. The red dashed line denotes perfect agreement between calculated and predicted energies, while the gray dashed lines mark the transition between negative and positive energies. Data points are colored according to their relative MD, as indicated by the color scale. Data points highlighted by red circles and labeled A−L correspond to selected outliers and are discussed in detail in Figure 4.

https://doi.org/10.1021/acs.jcim.5c03249
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

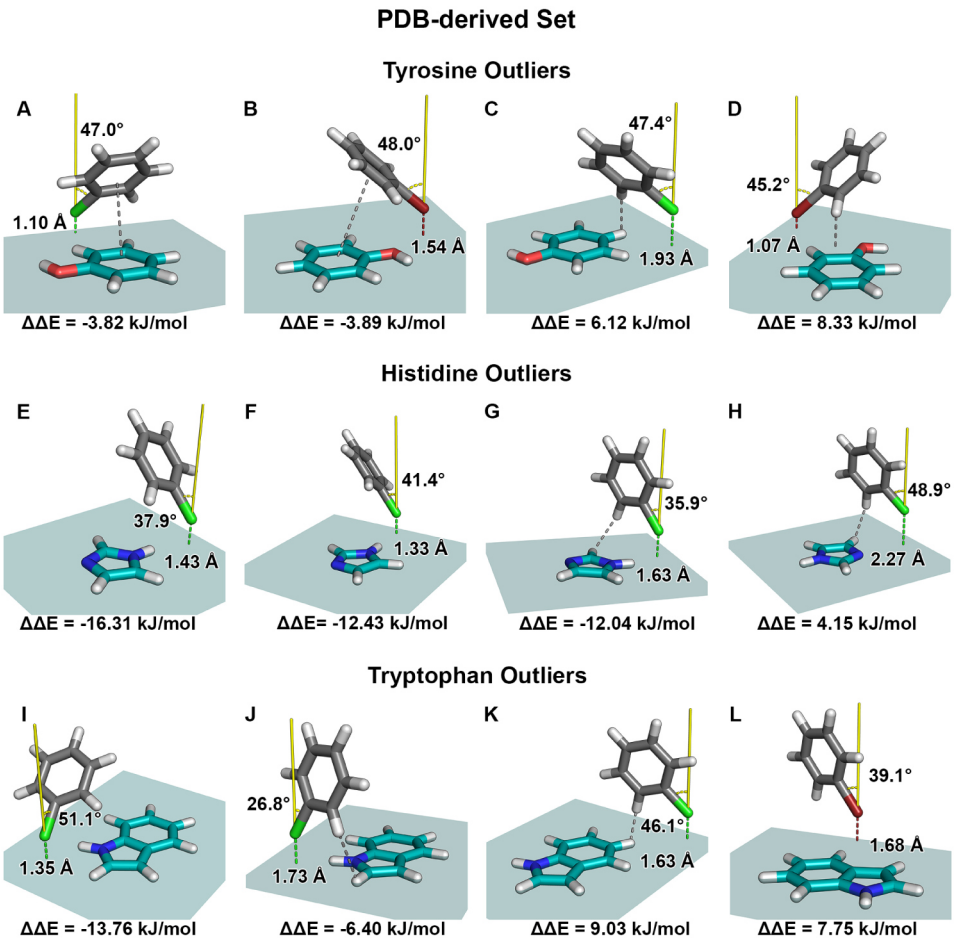


Figure 4. Interaction geometries for selected outliers from the PDB-derived geometry set involving the phenol, imidazole, and indole side chains of tyrosine, histidine, and tryptophan, as identified in the scatter plot shown in Figure 3. These structures correspond to data points exhibiting large deviations between the calculated and predicted adduct formation energies. Geometries A–D depict the relative orientation of the halobenzene (gray) with respect to the phenol π-system (teal). Shown are the halogen distance, $d_{X\cdots\pi\text{-plane}}$ (dashed line, colored according to the halogen), to the π-system plane in Å, the torsion angle, $\alpha_{C\text{-}X\cdots\perp(\pi\text{-plane})}$ (yellow), between the C-X bond vector and the normal to the π-system plane in degrees, and the corresponding energy difference, $\Delta\Delta E$. The gray dashed line indicates the type of interaction that may contribute to the observed prediction error. For clarity, the teal plane also illustrates the extent of the training grid used in the corresponding model. Geometries E–H show halobenzene interactions with imidazole, while geometries I–L show interactions with indole. Outliers of the tyrosine dataset: (A) Chlorobenzene interaction with $d_{Cl\cdots\pi\text{-plane}}$ = 1.10 Å and $\alpha_{C-Cl\cdots\perp(\pi\text{-plane})}$ = 47.0°. The gray dashed line indicates a $\pi\cdots\pi$-interaction. (B) Bromobenzene interaction with $d_{Br\cdots\pi\text{-plane}}$ = 1.54 Å and $\alpha_{C-Br\cdots\perp(\pi\text{-plane})}$ = 48.0°. The gray dashed line indicates a $\pi\cdots\pi$-interaction. (C) Chlorobenzene interaction with $d_{Cl\cdots\pi\text{-plane}}$ = 1.93 Å and $\alpha_{C-Cl\cdots\perp(\pi\text{-plane})}$ = 47.4°. The gray dashed line indicates a C–H$\cdots\pi$ contact. (D) Bromobenzene interaction with $d_{Br\cdots\pi\text{-plane}}$ = 1.07 Å and $\alpha_{C-Br\cdots\perp(\pi\text{-plane})}$ = 45.2°. The gray dashed line indicates a C–H$\cdots\pi$ contact. Outliers of the histidine dataset: (E) Chlorobenzene interaction with $d_{Cl\cdots\pi\text{-plane}}$ = 1.43 Å and $\alpha_{C-Cl\cdots\perp(\pi\text{-plane})}$ = 37.9°. (F) Chlorobenzene interaction with $d_{Cl\cdots\pi\text{-plane}}$ = 1.33 Å and $\alpha_{C-Cl\cdots\perp(\pi\text{-plane})}$ = 41.4°. (G) Chlorobenzene interaction with $d_{Cl\cdots\pi\text{-plane}}$ = 1.63 Å and $\alpha_{C-Cl\cdots\perp(\pi\text{-plane})}$ = 35.9°. The gray dashed line indicates a C–H$\cdots\pi$ contact. (H) Chlorobenzene interaction with $d_{Cl\cdots\pi\text{-plane}}$ = 2.27 Å and $\alpha_{C-Cl\cdots\perp(\pi\text{-plane})}$ = 48.9°. The gray dashed line indicates a C–H$\cdots\pi$ contact. Outliers of the tryptophan dataset: (I) Chlorobenzene interaction with $d_{Cl\cdots\pi\text{-plane}}$ = 1.35 Å and $\alpha_{C-Cl\cdots\perp(\pi\text{-plane})}$ = 51.1°. (J) Chlorobenzene interaction with $d_{Cl\cdots\pi\text{-plane}}$ = 1.73 Å and $\alpha_{C-Cl\cdots\perp(\pi\text{-plane})}$ = 26.8°. The gray dashed line indicates a $\pi\cdots\pi$ interaction. (K) Chlorobenzene interaction with $d_{Cl\cdots\pi\text{-plane}}$ = 1.63 Å and $\alpha_{C-Cl\cdots\perp(\pi\text{-plane})}$ = 46.1°. The gray dashed line indicates a C–H$\cdots\pi$ contact. (L) Bromobenzene interaction with $d_{Br\cdots\pi\text{-plane}}$ = 1.68 Å and $\alpha_{C-Br\cdots\perp(\pi\text{-plane})}$ = 39.1°.

These results confirm that the model interpolates reliably and is not dependent on memorization. In the full-data evaluation (step (ii) model performance decreased slightly but remained strong ($R^2_{TYR(ii)}$ = 0.9530, $RMSE_{TYR(ii)}$ = 0.76 kJ/mol). Although a few interactions exhibit larger discrepancies (up to +8.33 and −3.89 kJ/mol), these values show a steep improvement, where the positive deviations drop rapidly from +8.33 kJ/mol to +3 kJ/mol. Similarly, the negative tail of differences contracts below −2 kJ/mol after a few examples. Still, 20.23% lie beyond ±0.5 kJ/mol, but most structures remain accurately predicted (Figure 3, Tyrosine Set).

Representative outliers (Figure 4, examples A–D) exhibit large Mahalanobis distances, indicating that their geometric features lie at or beyond the boundaries of the training distribution, as previously described. Calculated and predicted adduct formation energies, energy differences, distances $d_{X\cdots\pi\text{-plane}}$ between the halogen and the π-plane, as well as the angle between the C-X vector and the normal to the π-plane $\alpha_{C\text{-}X\cdots\perp(\pi\text{-plane})}$ of all subsequent outliers are listed in Table 3. Examples A and B are dominated by $\pi\cdots\pi$-type arrangements characterized by strongly tilted halobenzene orientations and very short halogen$\cdots\pi$-plane separations. In example A, the exceptionally small distance of 1.10 Å places the halogen in

https://doi.org/10.1021/acs.jcim.5c03249
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Table 3. Overview of Energy Values and Geometric Parameters for Outlier Structures of the PDB-Derived Set of Tyrosine (Phenol, A−D), Histidine (Imidazole, E−H), and Tryptophan (Indole, I−L), Highlighted in Figure 3 and Depicted in Figure 4[a]

	halogen	addressed AA	PDB ID	ΔE_{calc} [kJ/mol]	ΔE_{pred} [kJ/mol]	$\Delta\Delta E$ [kJ/mol]	$d_{X\cdots\pi-plane}$ [Å]	$\alpha_{(C\text{-}X\cdots\perp(\pi-plane))}$ [°, deg]
A	Cl	TYR	6TOM	−15.08	−11.26	−3.82	1.1	46.97
B	Br	TYR	6FY4	5.45	9.34	−3.89	1.54	48.02
C	Cl	TYR	3UIC	−11.08	−17.19	6.12	1.93	47.45
D	Br	TYR	2WP5	−3.33	−11.66	8.33	1.07	45.17
E	Cl	HIS	3VAD	−19.16	−2.86	−16.31	1.43	37.92
F	Cl	HIS	9BST	−12.40	0.03	−12.43	1.33	41.36
G	Cl	HIS	2VPY	−8.70	3.34	−12.04	1.63	35.95
H	CL	HIS	6BYA	−3.02	−7.17	4.15	2.27	48.86
I	Cl	TRP	5LIK	−22.50	−8.74	−13.76	1.35	51.06
J	Cl	TRP	5IF2	−14.42	−8.02	−6.40	1.73	26.77
K	Cl	TRP	3DQT	−5.13	−14.16	9.03	1.63	46.09
L	Br	TRP	9EWA	9.19	1.44	7.75	1.68	39.13

[a]The halogen symbolizes the interacting halobenzene. The PDB ID indicates the crystal structure from which the interaction geometry was extracted. Values of calculated and predicted energies, as well as the difference between both, are given in kJ/mol. Distance values $d_{X\cdots\pi\text{-plane}}$ between the halogen and the π-plane are given in Å. Angle values between the C-X vector and the normal to the π-plane are given in degrees [°, deg].

close proximity to the aromatic system, enabling not only a favorable $\pi\cdots\pi$ interaction but also an additional stabilizing side interaction involving the halogen. Despite this, the interaction energy is underestimated by the model. In contrast, example B exhibits a similar tilt angle but a larger distance of 1.54 Å. Here, the halogen is positioned such that its negatively charged belt approaches the phenolic oxygen atom, leading to repulsive contributions that reduce the overall stabilization. It should be noted that the orientation of the hydroxyl group in these PDB-derived structures may differ from that in the optimized phenol reference configuration. A rigorous treatment of such effects would require more extensive and computationally demanding preprocessing.

Examples C and D are instead governed primarily by C−H$\cdots\pi$ motifs. In example C, the geometry represents an intermediate case between a C−H$\cdots\pi$ contact and a $\pi\cdots\pi$-like arrangement. Although the C−H positioning alone would suggest increased repulsion due to short H$\cdots$H contact, the combined interaction pattern remains stabilizing overall, as reflected by the calculated interaction energy of ΔE_{calc} = −11.08 kJ/mol. In example D, both a side-on H$\cdots$X interaction and a close approach of a halobenzene hydrogen toward the hydroxyl oxygen are observed. However, the overall geometry is excessively compact, leading to increased short-range repulsion and only weak net stabilization in the reference calculation (ΔE_{calc} = −3.33 kJ/mol). In contrast, the model predicts a much stronger attraction, indicating that repulsive contributions in such crowded geometries are not adequately captured.

Crucially, these interaction motifs fall outside the structural patterns characteristic of σ-hole contacts. The halogen frequently approaches from an off-axis position relative to the π-plane, or the halobenzene ring is positioned at unrealistically short distances, inducing either H$\cdots$X contacts or steric crowding and positive ΔE_{calc} values. Furthermore, the C−X bond is typically oriented away from the aromatic system, making the directional geometry required for σ-hole interactions unattainable.

Addressing the Side Chain of Histidine. The PDB-derived halogen$\cdots$HIS dataset contained 660 interactions. Of these, 128 structures satisfied the training-space restrictions (step (i)). Within this subset (denoted as "histidine set"), the imidazole model performed very well, achieving $R^2_{HIS(i)}$ = 0.9834 and $RMSE_{HIS(i)}$ = 0.43 kJ/mol, with deviations between +1.46 kJ/mol and −1.36 kJ/mol (21.09% beyond ±0.5 kJ/mol). Evaluation on the full His dataset (step (ii)) produced a moderate decline in accuracy ($R^2_{HIS(ii)}$ = 0.8840, $RMSE_{HIS(ii)}$ = 1.65 kJ/mol, 51.44% beyond ±0.5 kJ/mol; Figure S4). Outliers reached +4.15 kJ/mol and −16.31 kJ/mol, but the distribution shows similar behavior to the tyrosine case: positive deviations drop sharply after a few large positive values (to <2.5 kJ/mol), and only 18 structures exhibit negative deviations of more than −4 kJ/mol (Figure 3, Histidine Set).

Outlier analysis (Figure 4, examples E and H) confirms that these cases correspond to geometries far from the feature patterns represented during training, as reflected by high MDs. Individual energy, distance, and angle values are listed in Table 3. Many outliers display distorted orientations of the C-X bond relative to the imidazole π-system and comparatively short halogen$\cdots\pi$-plane separations. Across this series, the reference interaction energies become progressively less favorable from E to H. For examples E−G, the interaction energies are consistently underestimated by the model, whereas example H exhibits the opposite behavior and is overestimated. Notably, these geometries display smaller angle values than those observed for the phenol-derived outliers, yet substantial prediction errors persist. This indicates that the very short intermolecular distances alone are sufficient to place these structures outside the region of feature space sampled during training. Consequently, even when only one of the two key geometric descriptors—distance or angle—deviates strongly from the training distribution, the model struggles to reliably capture the balance between stabilizing and repulsive contributions. These results highlight a sensitivity of the imidazole model to compact C−H$\cdots\pi$ arrangements that are insufficiently represented in the training data.

Addressing the Side Chain of Tryptophan. 384 halogen$\cdots$TRP interactions were identified (denoted as the "tryptophan set"), of which 96 fell within the training distribution (step (i)). In this region, the indole model achieved excellent accuracy ($R^2_{TRP(i)}$ = 0.990, $RMSE_{TRP(i)}$ = 0.33 kJ/mol) with differences ranging between +1.18 kJ/mol and −0.38 kJ/mol (13.54% beyond ±0.5 kJ/mol). On the full dataset (step (ii)), performance decreased to $R^2_{TRP(ii)}$ =

https://doi.org/10.1021/acs.jcim.5c03249
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

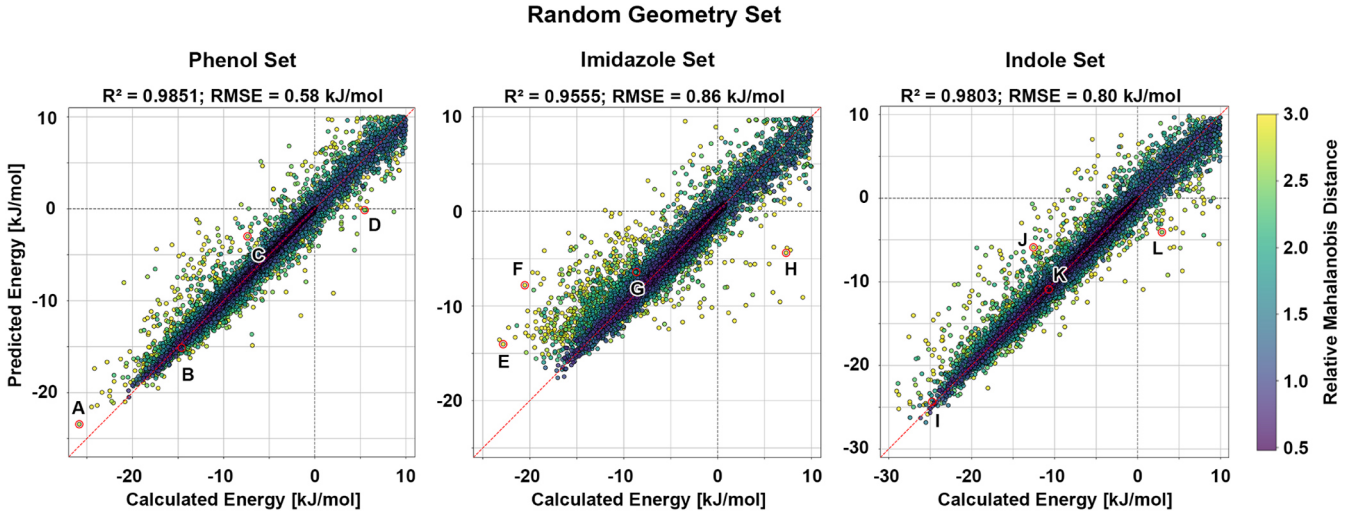


Figure 5. Model performance on the random geometry set after each model: Calculated adduct formation energies are plotted against the corresponding predicted values for the retrained phenol, imidazole, and indole models. The phenol model achieves an R^2 = 0.9851 with an RMSE = 0.58 kJ/mol, the imidazole model achieves an R^2 = 0.9555 with an RMSE = 0.86 kJ/mol, and the indole model achieves an R^2 = 0.9803 with an RMSE = 0.80 kJ/mol. The red dashed line denotes perfect agreement between calculated and predicted energies, while the gray dashed lines mark the transition between negative and positive energies. Data points are colored according to their relative MD, as indicated by the color scale. Data points highlighted by red circles and labeled A–L correspond to the previously discussed selected outliers (Figures 1 and 2).

0.8819 and $RMSE_{TRP(ii)}$ = 1.33 kJ/mol, with the largest deviations reaching +9.03 kJ/mol and −13.76 kJ/mol (28.65% beyond ±0.5 kJ/mol; Figure S4). As before, these values represent a small minority: positive deviations fall rapidly after the two largest examples to <2 kJ/mol, and only a few structures contribute to the most negative tail >−6 kJ/mol (Figure 3, Tryptophan Set).

The outliers (Figure 4, examples I–L) again correspond to geometries with elevated MDs, reflecting substantial deviation from the training distribution. These structures typically feature improper interaction angles, nonaxial halogen orientations, or sterically implausible proximities to the indole ring, thus displaying configurations inconsistent with physically meaningful σ-hole interactions. Examples I and J are characterized by underestimated interaction energies, with example I exhibiting a pronounced $\pi\cdots\pi$ interaction motif, while example J is dominated by a C–H$\cdots\pi$ contact. In the latter case, short H···H contacts would suggest an increased repulsive contribution. However, this effect is likely at least partially compensated for by a favorable H···X interaction, resulting in a net attractive reference energy. In contrast, examples K and L show overestimated interaction energies and are governed by weak C–H$\cdots\pi$ contacts or, in the case of example L, by the absence of a clearly identifiable stabilizing interaction motif.

Despite these differences, all four structures share very short halogen$\cdots\pi$-plane distances, while the corresponding interaction angles span a wide range.

Enhancing the Performance by Additional Training Data

Outlier analysis across all test sets and all three models indicates that poor predictive performance is consistently associated with elevated MD values. This relationship suggests that these poorly predicted interaction geometries possess features located at the boundaries of, or outside, the original training set distribution. These observations are consistent with our previous findings.[48] A commonly recognized advantage of neural network models is their modular expandability, both in terms of the learning architecture and the training data that can be incorporated during model development.

To evaluate this property in the present context, we retrained each model using an augmented training set that included an additional 300,000 randomly generated interaction geometries (100,000 per halogen, a similar procedure was used for the previous random geometry test sets). The underlying assumption is that broadening the training distribution enables the models to capture a wider range of structural features, thereby improving their ability to generalize to previously unseen or sparsely represented regions of feature space. A broader distribution should enhance the models' capacity to interpolate between data points and, consequently, reduce the error associated with examples that were previously identified as outliers.

All model configurations were kept unchanged, and each of the three systems—phenol, imidazole, and indole—was retrained using the expanded dataset. The resulting training performance remained high, with the retrained models yielding R^2 values of 0.9958, 0.9827, and 0.9952, and RMSE values of 0.34, 0.45, and 0.39 kJ/mol for phenol, imidazole, and indole, respectively. These values are slightly lower than those in the original training runs, which may be attributed to the broadened feature distribution: the newly introduced random geometries are inherently less systematically sampled and therefore comparatively underrepresented relative to the structured data present in the initial training set. As a result, the optimization process converges toward parameter sets that best describe the combined distribution on average, moderately reducing the apparent fit quality of the training data. Nevertheless, the central objective of this retraining step is to enhance the models' capacity for generalization, particularly for configurations previously identified as difficult to predict.

To assess this effect, the retrained models were first evaluated on the random geometry test set, exactly as used before. The updated performance metrics indicate improved predictive quality across all systems, with new R^2 values of 0.9851, 0.9555, and 0.9803, and RMSE values of 0.58, 0.86,

https://doi.org/10.1021/acs.jcim.5c03249
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

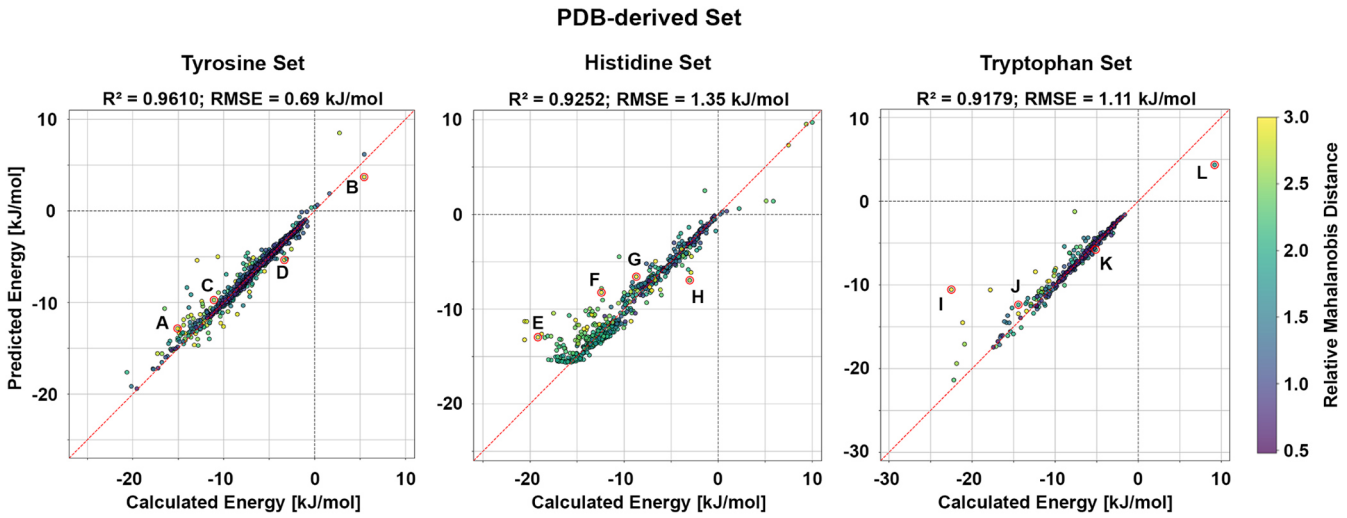


Figure 6. Model performance on the PDB-derived set after each model: Calculated adduct formation energies are plotted against the corresponding predicted values for the retrained phenol, imidazole, and indole models. The phenol model achieves an R^2_{TYR} = 0.9610 with an $RMSE_{TYR}$ = 0.69 kJ/mol, the imidazole model achieves an R^2_{HIS} = 0.9252 with an $RMSE_{HIS}$ = 1.35 kJ/mol, and the indole model achieves an R^2_{TRP} = 0.9179 with an $RMSE_{TRP}$ = 1.11 kJ/mol. The red dashed line denotes perfect agreement between calculated and predicted energies, while the gray dashed lines mark the transition between negative and positive energies. Data points are colored according to their relative MD, as indicated by the color scale. Data points highlighted by red circles and labeled A−L correspond to the previously discussed selected outliers (Figures 3 and 4).

and 0.80 kJ/mol for phenol, imidazole, and indole, respectively (Figure 5). These results demonstrate a pronounced reduction in prediction error for interaction geometries located near the boundaries of the original feature space, supporting the conclusion that the broadened training distribution enables more reliable interpolation in regions that were previously sparsely sampled. To illustrate the resulting improvement in predictive performance, the outlier geometries discussed above are highlighted in Figure 5. In contrast to the original models, these data points now cluster closely around the parity line, indicating a substantial enhancement in agreement between calculated and predicted energies.

To further validate the generalization behavior, the retrained models were subsequently reevaluated on the PDB-derived test set. Again, performance improvements were observed, with phenol, imidazole, and indole models achieving new R^2 values of 0.9610, 0.9252, and 0.9179, and corresponding RMSE values of 0.69, 1.35, and 1.11 kJ/mol. A complete overview of all R^2 and RMSE values of all models and evaluation steps of the different datasets is summarized in Table S3. Comparison with earlier results shows that the predicted energies now align more closely with the calculated reference values, with data points shifting noticeably toward the parity line in all three models (Figure 6). However, it is also evident that a subset of data points deviates further from the parity line than before. This behavior is consistent with the broader training distribution introduced during retraining. Because the model now optimizes its parameters to achieve the best average performance across a more diverse set of geometries, certain regions of the original systematic dataset become slightly less favored in the fitting process. These data points, therefore, incur a modest loss in accuracy as a consequence of the model's increased generalization. In principle, this effect could be mitigated by more carefully designing or weighting the additional training geometries, but such optimization and further refinement lie beyond the scope of the present study.

Taken together these findings indicate that expanding the training set with randomly generated interaction geometries not only increases model robustness but also significantly enhances the predictive accuracy for previously challenging regions of the interaction space.

CONCLUSION AND OUTLOOK

In the present work, we significantly extend our previous proof-of-concept study on neural network-based prediction of halogen···π interaction energies by moving from a single aromatic model system to a family of biologically relevant aromatic amino acid side chains. Dedicated models were developed for phenol, imidazole, and indole as representative models of tyrosine, histidine, and tryptophan, respectively. For each system, large, systematically generated training sets were constructed, and high-level MP2/TZVPP reference energies were used to train feed-forward neural networks on compact, rotationally and translationally invariant geometric descriptors. Across all three models, excellent predictive performance was achieved within the training domain, with R^2 values consistently above 0.98 and RMSE values well below 0.5 kJ/mol, demonstrating that the chosen feature representations and model architectures are well-suited to capture directional halogen···π interactions. Learning curve analysis further indicates that model performance converges rapidly with training set size, with near-optimal accuracy already achieved using roughly 30−40% of the available training data, while further data contribute to improved robustness.

A rigorous two-step evaluation strategy was employed to assess the generalization behavior both within and beyond the sampled feature space. While predictive accuracy remains high for geometries interpolating within the training distribution, systematic deviations were observed for compact or highly tilted configurations that lie outside the original training domain. Detailed outlier analysis revealed that these failures predominantly arise from interaction motifs not explicitly targeted during training, such as π···π or C−H···π contacts, or from short-range repulsive contributions that become dominant at very small intermolecular distances. Importantly, these limitations are consistent across all three aromatic systems and

https://doi.org/10.1021/acs.jcim.5c03249
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

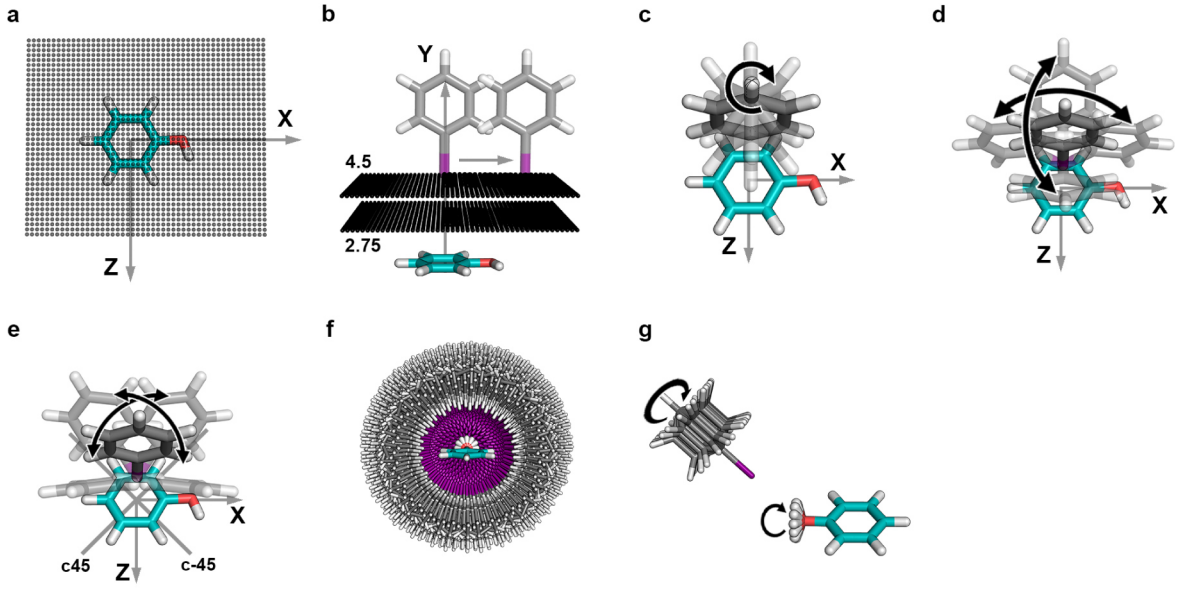


Figure 7. Illustration of the systematic generation of interaction geometries. Grid setups are depicted using phenol as a halogen bond acceptor. (a) Grid points on the XZ-plane were generated with dimensions $X_{translation}$ = [−5.0 Å to 7.0 Å], $Z_{translation}$ = [−5.0 Å to 5.0 Å] in steps of 0.25 Å. (b) Grid points were generated for eight different distances, $d_{X\cdots\pi\text{-plane}}$ = [2.75 to 4.5 Å] in steps of 0.25 Å, between the halogen atom (Cl, Br, or I) and the phenol plane. (c) Rotations of the halobenzene around the *y*-axis y_{rot} = [0° (initially), 45°, 90°, 135°]. (d) Deviations from the optimal *σ*-hole angle, $\alpha_{C\text{-}X\cdots\pi\text{-plane}}$ = 180°, from −40° to 40° in steps of 10°, achieved by rotating around the x- and *z*-axes. (e) Custom-generated rotational axes (c45 and c-45), lying 45° to the x- and *z*-axes. Rotations around these axes are similar to (d). (f) Halobenzenes addressing the phenol's oxygen atom were generated in a spherical shape around the oxygen with interaction distances $d_{X\cdots O}$ = [2.5−4.5 Å] in steps of 0.25 Å. The C-X vector points toward the oxygen atom, with $\alpha_{C\text{-}X\cdots O}$ = 180°. (g) Rotations of the halobenzene around the C-X axis (similar to (c)) in steps of 30°. Deviations of the phenol's O−H vector were made in steps of 30°. Figures were prepared with PyMOL.

mirror trends already observed in our earlier phenylalanine-based model, underscoring the generality of the underlying behavior rather than model-specific artifacts.

Application of the models to PDB-derived interaction geometries further demonstrates their practical relevance. For geometries consistent with *σ*-hole-driven halogen···*π* contacts, the models reproduce reference interaction energies with high accuracy, confirming their suitability for analyzing experimentally observed protein−ligand complexes. Deviations observed for certain PDB-derived structures can largely be traced back to geometries that violate the directional or distance requirements of genuine *σ*-hole interactions or to sterically crowded arrangements where repulsive effects dominate. These findings highlight the importance of combining geometric prefiltering with machine-learning-based energy evaluation when applying such models in real-world structural biology contexts.

A key advantage of this framework is its modular expandability. By augmenting the training data with additional randomly generated geometries, we demonstrated that model robustness and generalization can be substantially improved without alteration of the underlying architecture. The retrained models exhibit a distinct enhanced performance for previously challenging regions of feature space, including both random test geometries and PDB-derived geometries, albeit at the cost of a minor reduction in peak accuracy within the original, densely sampled domain. This trade-off illustrates a central design consideration for machine-learning approaches: balancing local precision with global transferability. As the halogen bond-specific model published recently by Devore and Shuford.[52] features interactions with ammonia as a standard acceptor, it unfortunately does not provide a valid comparison for our specialized model on halogen···*π* interaction energies. Thus, we turned to the more general-purpose machine-learned model (AIMNet2[53]), which was developed by Isayev and coworkers as a model that can, among many other applications, also recognize molecular interactions such as *σ*-hole interactions in halogen bonding. First results indicate that our highly specialized model is more suited for this particular application but is certainly limited to this specific task.

Taken together, this study establishes and confirms a scalable and transferable strategy for predicting halogen···*π* interaction energies across multiple aromatic environments with near-quantum-mechanical accuracy at negligible computational cost. By exploiting the "double jump" strategy from CCSD(T) to MP2 and ultimately to neural networks across several chemically distinct models, this work moves a decisive step closer toward practical deployment in structure-based drug design. Because the models are trained on gas-phase interaction energies, they are not intended to directly predict binding free energies. Instead, they are designed to complement existing docking and scoring frameworks, where solvation, desolvation penalties, and entropic effects are treated through implicit solvent models and empirical scoring terms. Future developments will focus on (i) incorporating representative repulsive interaction motifs into the training process to improve behavior at short range, (ii) extending the framework to positively charged *π*-systems such as protonated histidine (imidazolinium) and arginine (guanidinium), which represent a distinct class of halogen···*π* interaction motifs, (iii) extending the approach to additional noncovalent interaction types, and (iv) integrating the resulting models into molecular docking and scoring frameworks such as PLANTS, as well as web-based analysis platforms. Ultimately, this class of physically informed, residue-specific machine-learning models offers a promising route toward a more accurate and interpretable

https://doi.org/10.1021/acs.jcim.5c03249
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

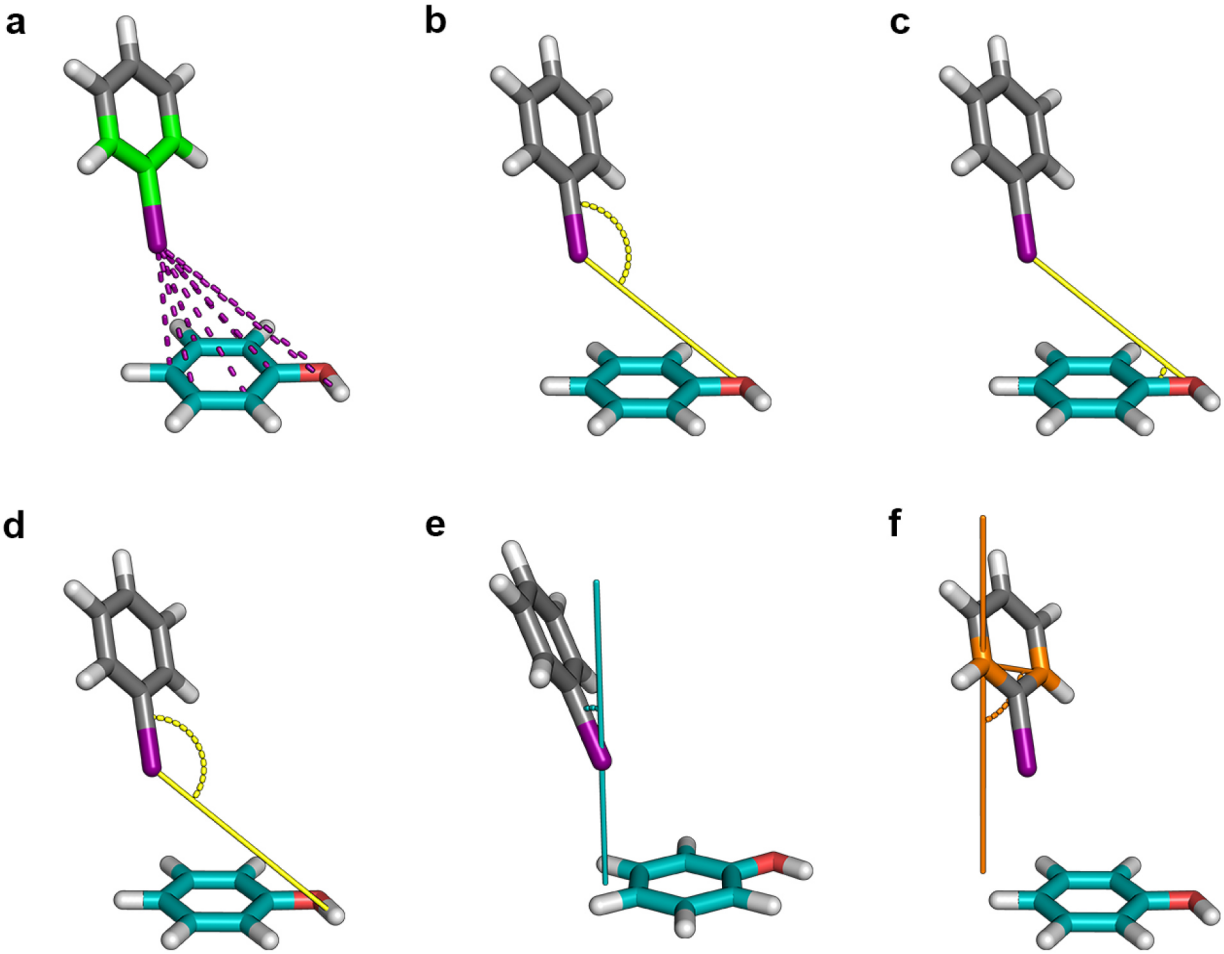


Figure 8. Overview of the feature extraction from halobenzene-phenol interaction geometries. (a) Pairwise distances from the halogen atom and the green-colored carbons of the halobenzene to all heavy atoms of the phenol system, as well as the hydrogen atom of the hydroxyl group. (b) Angle feature of the halogen atom, its neighboring carbon (C-X vector), and the phenol oxygen. (c) Angle feature of the phenol oxygen, its attached hydrogen, and the halogen atom. (d) Angle feature of the hydroxyl hydrogen and the C-X vector. (e) Angle feature of the normal vector of the phenol ring plane and the C-X vector. (f) Angle feature of the vector between the orange-colored carbons of the halobenzene and the normal of the phenol plane.

treatment of halogen bonding effects in biomolecular recognition.

MATERIALS AND METHODS

Structure Optimization

Geometry optimizations of the individual ligand model systems (iodobenzene, bromobenzene, and chlorobenzene) and the amino acid model systems (phenol, imidazole, and indole) were done at the MP2[54,55]-level of theory using TURBOMOLE 7.7.1[56] with a triple-ζ basis set (def2-TZVPP[57]) on the JUSTUS2 - bwHPC Cluster.[58] Calculations were done in combination with the resolution of identity (RI) technique and the frozen core approximation. Frozen core orbitals were defined using default settings, where orbitals with energies below −3.0 au are considered core orbitals. The SCF convergence criterion was increased to 10^{-8} hartree. Relativistic effects for iodine were considered by an effective core potential (ECP).[59−67]

Generation of Interaction Geometries

Interaction geometries of chloro-, bromo-, and iodobenzene in complex with phenol, imidazole, and indole were generated. To illustrate and describe the parameters of the geometry generation process, we subsequently used the complex of iodobenzene and phenol. Similar grids were generated for imidazole and indole interaction geometries with slightly adapted dimensions. The corresponding visualizations can be found in the Supporting Information (Figures S5 and S6). Halobenzenes were placed on a regular grid using X- and Z-translations (Figure 7a) for eight different distances (Figure 7b). Following previous approaches, an optimal σ-hole angle of $\alpha_{\text{C-X}\cdots\pi\text{-plane}} = 180°$ was initially used. To capture rotational features, halobenzene itself was rotated around the y-axis. Furthermore, the σ-hole angle was altered to deviations from −40° to 40° in steps of 10° in the x- and z-directions (Figure 7d). Two additional rotation axes were incorporated, lying at 45° to the x- or z-axis (Figure 7e). Additionally, interaction geometries addressing phenol's oxygen atom were generated in a spherical shape around the oxygen, where the C-X vector of the halobenzene points toward the oxygen atom with an angle of $\alpha_{\text{C-X}\cdots\text{O}} = 180°$ (Figure 7f). To capture different positions of the hydrogen atom attached to the oxygen, phenol's O−H vector was altered in steps of 30° (Figure 7g). For imidazole complexes, additional interaction geometries addressing the nonprotonated nitrogen atom were generated in a spherical shape around the nitrogen, where the C-X vector of the halobenzene points toward the nitrogen atom with an angle of $\alpha_{\text{C-X}\cdots\text{N}} = 180°$. Single-point calculations of more than 18.6 million (total sum of the three systems) interaction geometries were conducted at the MP2/TZVPP level of theory. Adduct formation energies were calculated as:

$$\Delta E = (E_{\text{complex}} - (E_{\text{halobenzene}} + E_{\text{phenol/indole}})) \quad (1)$$

and reported as kJ/mol.

https://doi.org/10.1021/acs.jcim.5c03249
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

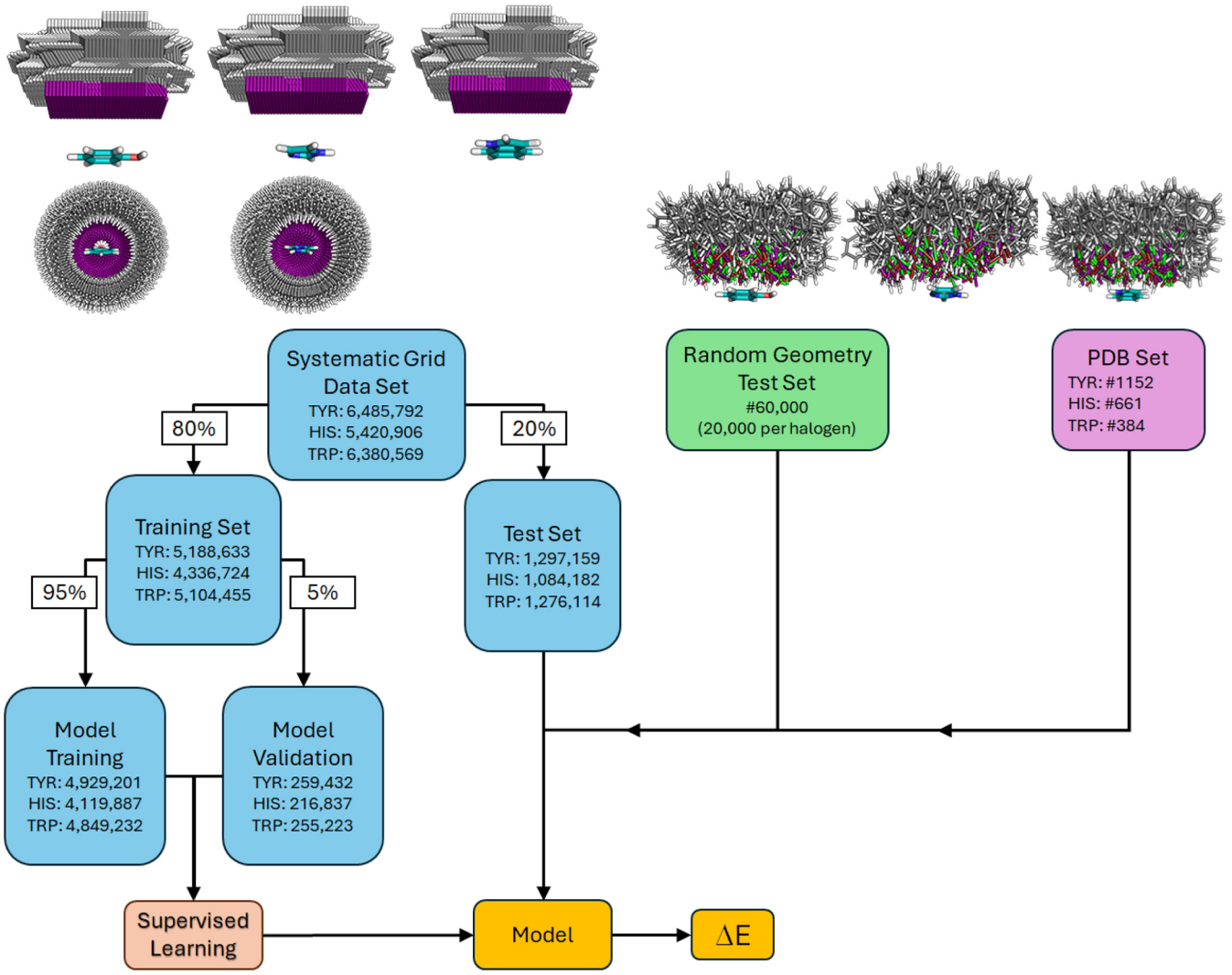


Figure 9. Overview of the different datasets. The systematic grid dataset is split into a training set (80% of the whole dataset) and a test set (20%). The training set is further split into model training (again 80%) and the model validation set (20%). The two model datasets are fed into the model via a supervised training approach. The random geometry test set (60,000 geometries) is used to evaluate the models' generalized performance on unseen data. The PDB set (N_{TYR} = 1152, N_{HIS} = 661, and N_{TRP} = 384) is used to represent and evaluate biological examples. The respective geometries shown are only a small excerpt of the full datasets.

Generation of a Random Geometry Training and Test Set

Similar to the generation of the systematic dataset, we generated a smaller subset of 60,000 interaction geometries (20,000 per halobenzene) with random geometric features for each addressed system to test the models' generalizability. Parameters of X, $Z_{translation}$ = [−5.0 to 7.0 Å], $Y_{translation}$ = [1.5 to 5.0 Å], y_{rot} = [0° to 360°], and $\alpha_{C\text{-}X\cdots\perp(\pi\text{-plane})}$ = [0° to 60°] were randomly chosen and applied to a halobenzene. To ensure the uniqueness of the generated geometries, newly chosen parameters were compared to previous ones and only applied if found to be distinct (>0.4 Å between each halogen atom and each neighboring carbon atom of two given molecules) within the dataset. Datasets for retraining (300,000 geometries, 100,000 per halogen) were similarly generated for each model system.

Feature Extraction and ANN Model Training

All data preprocessing, feature extraction, and learning approaches for all three model systems were built in Python using custom scripts with the *PyTorch*[68] and *scikit-learn* packages, two open-source Python libraries for machine learning. ANN models were trained on geometric features extracted from the interaction geometries. The training process for the individual models was performed on the BinAC—bwHPC Cluster.[69] In general, the feature vectors $\vec{v} = (d_1, d_2, \ldots, d_n, a_1, \ldots, a_n)$ comprise individual distance and angle descriptors, with all features detailed in Tables S1 and S2. The adduct formation energy ΔE of each geometry serves as the target value. Distance features were computed as all pairwise distances between selected atoms of the halobenzene fragment (the halogen atom, its neighboring carbon, and the two adjacent ring carbons) and every heavy atom of the respective amino acid model system (Figure 8a). Additional distances involving the phenol hydroxyl hydrogen and the imidazole N−H hydrogen were also included.

Angle features were derived from vectors defined between halobenzene and the amino acid model system. For phenol, five angle descriptors were used: (i) the angle between the halogen atom, its neighboring carbon (CX vector), and the phenol oxygen (Figure 8b); (ii) the angle formed by the phenol oxygen, its attached hydrogen, and the halogen atom (Figure 8c); (iii) the angle between the hydroxyl hydrogen and the C-X vector (Figure 8d); (iv) the angle between the normal vector of the phenol ring plane and the C-X vector (Figure 8e); and (v) the angle between the normal vector of the phenol ring plane and the vector defined by the two adjacent halobenzene ring carbons (Figure 8f). The imidazole feature set incorporates four angle descriptors (Figure S5): (i) the angle between the nonprotonated nitrogen and the C-X vector; (ii) the angle between the protonated nitrogen and the C-X vector; (iii) the angle between the normal of the imidazole plane and the C-X vector; and

https://doi.org/10.1021/acs.jcim.5c03249
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

(iv) the angle between the vector defined by the two adjacent halobenzene ring carbons and the normal of the imidazole plane. The indole feature vector contains three angle descriptors (Figure S6): (i) the angle between the C-X vector and the center of mass of the indole heterocycle; (ii) the angle between the indole-plane normal and the C-X vector; and (iii) the angle between the vector defined by the two adjacent halobenzene ring carbons and the indole-plane normal. In total, the phenol, imidazole, and indole feature vectors comprise 37, 28, and 39 distance and angle features, respectively.

Prior to model training, each feature was scaled independently using a min-max normalization procedure (scikit-learn MinMaxScaler). The complete dataset was partitioned into an 80% training portion and a 20% test portion using a stratified 5-fold leave-one-out scheme to ensure reproducibility and balanced sampling (Figure 9). Stratification was performed with respect to (i) the halogen identity and (ii) the halogen···π-plane distance, thereby maintaining comparable feature distributions across all folds. Each training fold was subsequently divided again into a training subset (80%) and an internal validation subset (20%) by using the same stratification criteria.

Model development involves extensive hyperparameter optimization within a supervised learning framework. The search space included different activation functions (Sigmoid, Tanh, and LeakyReLU from PyTorch), network depths of two to four fully connected hidden layers, and a variety of layer-width combinations drawn from [256, 128, 64, 32, 16, and 8]. Additional hyperparameters, including initial learning rate (0.1 to 0.0001), batch size (32 to 2048), and number of training epochs (100 to 10 000), were systematically varied. Training was performed using the Adam optimizer (PyTorch) and an elastic-net-weighted MSE loss to mitigate data imbalance. Early stopping criteria were employed to avoid overfitting. The hyperparameter set that performed best on the validation subset was used for a final training run on a 95%/5% training/validation set split. Model accuracy was assessed using the mean squared error (MSE), root-mean-square error (RMSE), and coefficient of determination (R^2). High predictive quality is reflected by an R^2 value approaching 1.0 in combination with low MSE and RMSE values. The RMSE per epoch from the final training phase is presented in Figure S2a,c, and e for each system individually. Energy deviations between computed and predicted adduct formation energies were calculated as in kJ/mol.

$$\Delta\Delta E = \Delta E_{\text{calculated}} - \Delta E_{\text{predicted}} \tag{2}$$

The training started with a fixed learning rate and was adapted (the value was halved) down to a minimum of 0.0001 if there had not been a loss improvement for 10 epochs. The final hyperparameter configuration for all three models comprised three fully connected hidden layers with 256, 128, and 64 units, respectively, using Leaky Rectified Linear Unit (Leaky ReLU) activation functions. The initial learning rate was set to 0.001, while batch sizes were set to 1024 for the imidazole model and 2048 for the phenol and indole models.

PDB Scan of Tyrosine, Histidine, and Tryptophan Acceptors

A PDB (as of July 2025:239, 149) scan was conducted using a custom Python/PyMOL[70] script. All protein structures were preprocessed by removing alternative conformations, metal ions, and hydrogen atoms. PDB entries were then screened for ligands containing aromatic halogens (Cl, Br, or I) bound to an aromatic ring and comprising at least six heavy atoms. Aromatic amino acid side chain residues of tyrosine, histidine, and tryptophan within 5 Å of the halogen atom were retained as XB acceptors. To increase the chance for plausible XB geometries, an angular criterion was applied, where the C-X bond vector was required to orient toward the π-plane of the residue ($\alpha_{\text{C-X}\cdots\perp(\pi\text{-plane})} < 50°$). A lower distance cutoff of $d_{\text{X}\cdots\text{AA}} = 1$ Å was employed. The side chain residue of each addressed amino acid was replaced by the respective model system (phenol for tyrosine, imidazole for histidine, and indole for tryptophan). Likewise, the native ligand was substituted by the matching halobenzene (chloro-, bromo-, or iodobenzene), ensuring that the halogen atom position, the C-X bond direction, and the orientation of the aromatic or heteroaromatic plane were preserved by following a matched molecular pair strategy. Both molecular fragments were separately optimized at the MP2/TZVPP level of theory, as described previously. From this procedure, we obtained $N_{\text{TYR}} = 1152$, $N_{\text{HIS}} = 661$, and $N_{\text{TRP}} = 384$ distinct interaction geometries between halobenzene and the respective model system. Single-point calculations were subsequently carried out at the MP2/TZVPP level. Adduct formation energies were computed by using the supermolecular scheme (eq 1).

Model Evaluation and Outlier Detection

To assess how well the models reproduce adduct formation energies for a given interaction geometry, three independent test sets were evaluated: the 20% withheld test set, the random-geometry test set, and the PDB-derived set. Ideally, these test data represent previously unseen configurations, allowing a meaningful assessment of the models' generalizability. Model performance was quantified using the root-mean-square error (RMSE) and coefficient of determination (R^2). Because the random-geometry and PDB test sets may contain structures that fall outside the feature domain sampled during training, we additionally characterized the degree of extrapolation using the Mahalanobis distance (MD). MD provides a multivariate measure of how far a data point lies from the centroid of the training set feature distribution. It is defined as:

$$D_M(x) = \sqrt{(x-\mu)^T \Sigma^{-1} (x-\mu)} \tag{3}$$

where x denotes the feature vector of the input geometry, μ is the mean feature vector of the training data, and Σ is the corresponding covariance matrix. Large MD values indicate that a geometry occupies a low-probability region of feature space relative to the training data, marking it as a potential outlier or extrapolative case. MDs were calculated for all test set geometries, and a threshold was established to identify points classified as outliers.

■ ASSOCIATED CONTENT

Data Availability Statement

PyMOL is an open-source software maintained and distributed by Schrödinger. There is an open-source version of PyMOL available at: https://github.com/schrodinger/pymol-open-source. Python and all of its' packages are an open-source programming language available and downloadable from https://www.python.org/. PyTorch is an open-source machine learning library for Python: https://pytorch.org/. TURBOMOLE is a purchasable software maintained and distributed by *TURBOMOLE GmbH*. Demo versions are available at https://www.turbomole.org/. The licensed software was provided to us by the bwHPC Cluster JUSTUS2.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.5c03249.

Graphical depiction of distance- and angle-based features; detailed description of the features used for training; results of the training and validation process of final models; histograms of energy differences and random geometry and PDB-derived test sets; graphical depiction of grid generation and feature extraction for imidazole and indole datasets; summary of model performances; and learning curve analysis (PDF)

Detailed tables of the individual data points of the random geometry dataset and the PDB dataset with corresponding adduct formation energies for calculated and predicted energies (XLSX)

https://doi.org/10.1021/acs.jcim.5c03249
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

AUTHOR INFORMATION

Corresponding Author

Frank M. Boeckler − *Laboratory for Molecular Design & Pharmaceutical Biophysics, Institute of Pharmaceutical Sciences, Department of Pharmacy and Biochemistry, Eberhard Karls Universität Tübingen, Tübingen 72076, Germany; Interfaculty Institute for Biomedical Informatics (IBMI), Eberhard Karls Universität Tübingen, Tübingen 72076, Germany;* orcid.org/0000-0001-8738-6716; Email: frank.boeckler@uni-tuebingen.de

Authors

Marc U. Engelhardt − *Laboratory for Molecular Design & Pharmaceutical Biophysics, Institute of Pharmaceutical Sciences, Department of Pharmacy and Biochemistry, Eberhard Karls Universität Tübingen, Tübingen 72076, Germany;* orcid.org/0009-0007-9152-8538

Finn Mier − *Laboratory for Molecular Design & Pharmaceutical Biophysics, Institute of Pharmaceutical Sciences, Department of Pharmacy and Biochemistry, Eberhard Karls Universität Tübingen, Tübingen 72076, Germany;* orcid.org/0000-0003-3409-4626

Markus O. Zimmermann − *Laboratory for Molecular Design & Pharmaceutical Biophysics, Institute of Pharmaceutical Sciences, Department of Pharmacy and Biochemistry, Eberhard Karls Universität Tübingen, Tübingen 72076, Germany; Interfaculty Institute for Biomedical Informatics (IBMI), Eberhard Karls Universität Tübingen, Tübingen 72076, Germany;* orcid.org/0000-0001-6115-8248

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.5c03249

Author Contributions

The manuscript was written through the contributions of all authors. All authors have given their approval to the final version of the manuscript. F.M.B. and M.U.E. envisioned the research. M.U.E. conducted the majority of the QM calculations, developed all machine learning and evaluation scripts, assembled the complete dataset of results, and prepared all visualizations. M.O.Z. performed a small subset of the QM calculations. M.U.E. prepared the original draft. F.M. contributed to developing the computational strategy and provided comments on the manuscript. M.U.E., F.M., M.O.Z., and F.M.B. reviewed, edited, and finalized the manuscript.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors acknowledge support from the state of Baden-Württemberg through bwHPC and the German Research Foundation (DFG) through Grant No. INST 40/575-1 FUGG (JUSTUS 2 cluster). In addition, support is acknowledged from the High Performance and Cloud Computing Group at the Zentrum für Datenverarbeitung of the University of Tübingen and the German Research Foundation (DFG) through Grant No. INST 37/935-1 FUGG (BinAC cluster).

ABBREVIATIONS

XB, halogen bond; X, halogen; QM, quantum mechanical; MP2, Møller−Plesset perturbation method order 2; TZVPP, valence triple-ζ with two sets of polarization functions; SCF, self-consistent field; RI, resolution of identity; RMSE, root-mean-square error; ECP, effective core potential; NN, neural network; R^2, coefficient of determination; MD, Mahalanobis Distance; PDB, Protein Data Bank; PHE, phenylalanine; TYR, tyrosine; HIS, histidine; TRP, tryptophan; MSE, mean squared error; CCSD(T), coupled cluster with single, double, and perturbative triple excitations

REFERENCES

(1) Bissantz, C.; Kuhn, B.; Stahl, M. A Medicinal Chemist's Guide to Molecular Interactions. *J. Med. Chem.* **2010**, *53* (14), 5061−5084.

(2) Adhav, V. A.; Saikrishnan, K. The Realm of Unconventional Noncovalent Interactions in Proteins: Their Significance in Structure and Function. *ACS Omega* **2023**, *8* (25), 22268−22284.

(3) Anighoro, A. Underappreciated Chemical Interactions in Protein−Ligand Complexes. *Quantum Mech. Drug Discovery* **2020**, 75−86.

(4) Jena, S.; Dutta, J.; Tulsiyan, K. D.; Sahu, A. K.; Choudhury, S. S.; Biswal, H. S. Noncovalent interactions in proteins and nucleic acids: Beyond hydrogen bonding and π-stacking. *Chem. Soc. Rev.* **2022**, *51* (11), 4261−4286.

(5) Müller-Dethlefs, K.; Hobza, P. Noncovalent Interactions: A Challenge for Experiment and Theory. *Chem. Rev.* **2000**, *100* (1), 143−168.

(6) Clark, T.; Hennemann, M.; Murray, J. S.; Politzer, P. Halogen bonding: The σ-hole. *J. Mol. Model.* **2007**, *13* (2), 291−296.

(7) Desiraju, G. R.; Ho, P. S.; Kloo, L.; Legon, A. C.; Marquardt, R.; Metrangolo, P.; Politzer, P.; Resnati, G.; Rissanen, K. Definition of the halogen bond (IUPAC Recommendations 2013). *Pure Appl. Chem.* **2013**, *85* (8), 1711−1713.

(8) Politzer, P.; Murray, J. S.; Clark, T. Halogen bonding: An electrostatically-driven highly directional noncovalent interaction. *Phys. Chem. Chem. Phys.* **2010**, *12* (28), 7748−7757.

(9) Wang, C.; Danovich, D.; Mo, Y.; Shaik, S. On The Nature of the Halogen Bond. *J. Chem. Theory Comput.* **2014**, *10* (9), 3726−3737.

(10) Cavallo, G.; Metrangolo, P.; Milani, R.; Pilati, T.; Priimagi, A.; Resnati, G.; Terraneo, G. The Halogen Bond. *Chem. Rev.* **2016**, *116* (4), 2478−2601.

(11) Heidrich, J.; Exner, T. E.; Boeckler, F. M. Predicting the Magnitude of σ-Holes Using VmaxPred, a Fast and Efficient Tool Supporting the Application of Halogen Bonds in Drug Discovery. *J. Chem. Inf. Model.* **2019**, *59* (2), 636−643.

(12) Bhattarai, S.; Sutradhar, D.; Chandra, A. K. Tuning of halogen-bond strength: Comparative role of basicity and strength of σ-hole. *J. Mol. Struct.* **2021**, *1223*, 129239.

(13) Donald, K. J.; Pham, N.; Ravichandran, P. Sigma Hole Potentials as Tools: Quantifying and Partitioning Substituent Effects. *J. Phys. Chem. A* **2023**, *127* (48), 10147−10158.

(14) Esrafili, M. D.; Mahdavinia, G.; Javaheri, M.; Sobhi, H. R. A theoretical study of substitution effects on halogen−π interactions. *Mol. Phys.* **2014**, *112* (8), 1160−1166.

(15) Lange, A.; Heidrich, J.; Zimmermann, M. O.; Exner, T. E.; Boeckler, F. M. Scaffold Effects on Halogen Bonding Strength. *J. Chem. Inf. Model.* **2019**, *59* (2), 885−894.

(16) Sedlak, R.; Kolář, M. H.; Hobza, P. Polar Flattening and the Strength of Halogen Bonding. *J. Chem. Theory Comput.* **2015**, *11* (10), 4727−4732.

(17) Berger, G.; Frangville, P.; Meyer, F. Halogen bonding for molecular recognition: New developments in materials and biological sciences. *Chem. Commun.* **2020**, *56* (37), 4970−4981.

(18) Erdélyi, M. Halogen bonding in solution. *Chem. Soc. Rev.* **2012**, *41* (9), 3547−3557.

(19) Hardegger, L. A.; Kuhn, B.; Spinnler, B.; Anselm, L.; Ecabert, R.; Stihle, M.; Gsell, B.; Thoma, R.; Diez, J.; Benz, J.; et al. Systematic Investigation of Halogen Bonding in Protein−Ligand Interactions. *Angew. Chem., Int. Ed.* **2011**, *50* (1), 314−318.

(20) Hernandes, Z. M.; Cavalcanti, T. S. M.; Moreira, M. D. R.; de Azevedo Junior, F. W.; Leite, L. A. C. Halogen Atoms in the Modern

Medicinal Chemistry: Hints for the Drug Design. *Curr. Drug Targets* **2010**, *11* (3), 303−314.
(21) Parker, A. J.; Stewart, J.; Donald, K. J.; Parish, C. A. Halogen Bonding in DNA Base Pairs. *J. Am. Chem. Soc.* **2012**, *134* (11), 5165−5172.
(22) Scholfield, M. R.; Zanden, C. M. V.; Carter, M.; Ho, P. S. Halogen bonding (X-bonding): A biological perspective. *Protein Sci.* **2013**, *22* (2), 139−152.
(23) Shinada, N. K.; de Brevern, A. G.; Schmidtke, P. Halogens in Protein−Ligand Binding Mechanism: A Structural Perspective. *J. Med. Chem.* **2019**, *62* (21), 9341−9356.
(24) Sirimulla, S.; Bailey, J. B.; Vegesna, R.; Narayan, M. Halogen Interactions in Protein−Ligand Complexes: Implications of Halogen Bonding for Rational Drug Design. *J. Chem. Inf. Model.* **2013**, *53* (11), 2781−2791.
(25) Xu, Z.; Yang, Z.; Liu, Y.; Lu, Y.; Chen, K.; Zhu, W. Halogen Bond Its Role beyond Drug−Target Binding Affinity for Drug Discovery and Development. *J. Chem. Inf. Model.* **2014**, *54* (1), 69−78.
(26) Walker, M. G.; Mendez, C. G.; Ho, A. N.; Czarny, R. S.; Rappé, A. K.; Ho, P. S. Design of a halogen bond catalyzed DNA endonuclease. *Proc. Natl. Acad. Sci. U. S. A.* **2025**, *122* (14), No. e2500099122.
(27) Wilcken, R.; Zimmermann, M. O.; Bauer, M. R.; Rutherford, T. J.; Fersht, A. R.; Joerger, A. C.; Boeckler, F. M. Experimental and Theoretical Evaluation of the Ethynyl Moiety as a Halogen Bioisostere. *ACS Chem. Biol.* **2015**, *10* (12), 2725−2732.
(28) Dammann, M.; Stahlecker, J.; Zimmermann, M. O.; Klett, T.; Rotzinger, K.; Kramer, M.; Coles, M.; Stehle, T.; Boeckler, F. M. Screening of a Halogen-Enriched Fragment Library Leads to Unconventional Binding Modes. *J. Med. Chem.* **2022**, *65* (21), 14539−14552.
(29) Vaas, S.; Zimmermann, M. O.; Schollmeyer, D.; Stahlecker, J.; Engelhardt, M. U.; Rheinganz, J.; Drotleff, B.; Olfert, M.; Lämmerhofer, M.; Kramer, M.; et al. Principles and Applications of CF2X Moieties as Unconventional Halogen Bond Donors in Medicinal Chemistry, Chemical Biology, and Drug Discovery. *J. Med. Chem.* **2023**, *66* (15), 10202−10225.
(30) Jiang, L.; Zhang, X.; Zhou, Y.; Chen, Y.; Luo, Z.; Li, J.; Yuan, C.; Huang, M. Halogen bonding for the design of inhibitors by targeting the S1 pocket of serine proteases. *RSC Adv.* **2018**, *8* (49), 28189−28197.
(31) Lu, Y.; Liu, Y.; Li, H.; Zhu, X.; Liu, H.; Zhu, W. Energetic Effects between Halogen Bonds and Anion-π or Lone Pair-π Interactions: A Theoretical Study. *J. Phys. Chem. A* **2012**, *116* (10), 2591−2597.
(32) Matter, H.; Nazaré, M.; Güssregen, S.; Will, D. W.; Schreuder, H.; Bauer, A.; Urmann, M.; Ritter, K.; Wagner, M.; Wehner, V. Evidence for C-Cl/C-Br$\cdots\pi$ Interactions as an Important Contribution to Protein−Ligand Binding Affinity. *Angew. Chem., Int. Ed.* **2009**, *48* (16), 2911−2916.
(33) Parisini, E.; Metrangolo, P.; Pilati, T.; Resnati, G.; Terraneo, G. Halogen bonding in halocarbon−protein complexes: A structural survey. *Chem. Soc. Rev.* **2011**, *40* (5), 2267−2278.
(34) Politzer, P.; Murray, J. S.; Clark, T. Halogen bonding and other σ-hole interactions: A perspective. *Phys. Chem. Chem. Phys.* **2013**, *15* (27), 11178−11189.
(35) Rowe, R. K.; Ho, P. S. Relationships between hydrogen bonds and halogen bonds in biological systems. *Acta Crystallogr., Sect. B* **2017**, *73* (2), 255−264.
(36) Ford, M. C.; Ho, P. S. Computational Tools To Model Halogen Bonds in Medicinal Chemistry. *J. Med. Chem.* **2016**, *59* (5), 1655−1670.
(37) Young, D. C. *Computational Chemistry: A Practical Guide for Applying Techniques to Real World Problems*; Wiley, 2001. DOI: .
(38) Zhu, Z.; Xu, Z.; Zhu, W. Interaction Nature and Computational Methods for Halogen Bonding: A Perspective. *J. Chem. Inf. Model.* **2020**, *60* (6), 2683−2696.
(39) Wilcken, R.; Zimmermann, M. O.; Lange, A.; Zahn, S.; Boeckler, F. M. Using halogen bonds to address the protein backbone: A systematic evaluation. *J. Comput. Aided Mol. Des.* **2012**, *26* (8), 935−945.
(40) Wilcken, R.; Zimmermann, M. O.; Lange, A.; Joerger, A. C.; Boeckler, F. M. Principles and Applications of Halogen Bonding in Medicinal Chemistry and Chemical Biology. *J. Med. Chem.* **2013**, *56* (4), 1363−1388.
(41) Zimmermann, M. O.; Boeckler, F. M. Targeting the protein backbone with aryl halides: Systematic comparison of halogen bonding and $\pi\cdots\pi$ interactions using N-methylacetamide. *Med. Chem. Commun.* **2016**, *7* (3), 500−505.
(42) Wilcken, R.; Zimmermann, M. O.; Lange, A.; Zahn, S.; Kirchner, B.; Boeckler, F. M. Addressing Methionine in Molecular Design through Directed Sulfur−Halogen Bonds. *J. Chem. Theory Comput.* **2011**, *7* (7), 2307−2315.
(43) Lange, A.; Zimmermann, M. O.; Wilcken, R.; Zahn, S.; Boeckler, F. M. Targeting Histidine Side Chains in Molecular Design through Nitrogen−Halogen Bonds. *J. Chem. Inf. Model.* **2013**, *53* (12), 3178−3189.
(44) Zimmermann, M. O.; Lange, A.; Zahn, S.; Exner, T. E.; Boeckler, F. M. Using Surface Scans for the Evaluation of Halogen Bonds toward the Side Chains of Aspartate, Asparagine, Glutamate, and Glutamine. *J. Chem. Inf. Model.* **2016**, *56* (7), 1373−1383.
(45) Engelhardt, M. U.; Zimmermann, M. O.; Dammann, M.; Stahlecker, J.; Poso, A.; Kronenberger, T.; Kunick, C.; Stehle, T.; Boeckler, F. M. Halogen Bonding on Water—A Drop in the Ocean? *J. Chem. Theory Comput.* **2024**, *20* (23), 10716−10730.
(46) Zhou, P.; Lv, J.; Zou, J.; Tian, F.; Shang, Z. Halogen−water−hydrogen bridges in biomolecules. *J. Struct Biol.* **2010**, *169* (2), 172−182.
(47) Engelhardt, M. U.; Zimmermann, M. O.; Mier, F.; Boeckler, F. M. Comparison of QM Methods for the Evaluation of Halogen−π Interactions for Large-Scale Data Generation. *J. Chem. Theory Comput.* **2025**, *21* (12), 6174−6183.
(48) Engelhardt, M. U.; Mier, F.; Zimmermann, M. O.; Boeckler, F. M. A QM-AI Approach for the Acceleration of Accurate Assessments of Halogen-π Interactions by Training Neural Networks. *J. Chem. Inf. Model.* **2025**, *65*, 13132.
(49) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28* (1), 235−242.
(50) Mahalanobis, P. C.On the generalized distance in statistics*Proceedings of the National Institute of Sciences of India*National Institute of Sciences of India19361249−55
(51) Perez, M. A. S.; Bassani-Sternberg, M.; Coukos, G.; Gfeller, D.; Zoete, V. Analysis of Secondary Structure Biases in Naturally Presented HLA-I Ligands. *Front. Immunol.* **2019**, *10*, 2731.
(52) Devore, D. P.; Shuford, K. L. Data and Molecular Fingerprint-Driven Machine Learning Approaches to Halogen Bonding. *J. Chem. Inf. Model.* **2024**, *64* (21), 8201−8214.
(53) Anstine, D. M.; Zubatyuk, R.; Isayev, O. AIMNet2: A neural network potential to meet your neutral, charged, organic, and elemental-organic needs. *Chem. Sci.* **2025**, *16* (23), 10228−10244.
(54) Head-Gordon, M.; Pople, J. A.; Frisch, M. J. MP2 energy evaluation by direct methods. *Chem. Phys. Lett.* **1988**, *153* (6), 503−506.
(55) Møller, C.; Plesset, M. S. Note on an Approximation Treatment for Many-Electron Systems. *Phys. Rev.* **1934**, *46* (7), 618−622.
(56) *TURBOMOLE V7.7.1 2019*; A development of University of Karlsruhe and Forschungszentrum Karlsruhe GmbH, 1989−2007, TURBOMOLE GmbH, 2007.
(57) Schäfer, A.; Huber, C.; Ahlrichs, R. Fully optimized contracted Gaussian basis sets of triple zeta valence quality for atoms Li to Kr. *J. Chem. Phys.* **1994**, *100* (8), 5829−5835.
(58) *bwForCluster - JUSTUS 2*; https://wiki.bwhpc.de/e/JUSTUS2.
(59) Feyereisen, M.; Fitzgerald, G.; Komornicki, A. Use of approximate integrals in ab initio theory. An application in MP2 energy calculations. *Chem. Phys. Lett.* **1993**, *208* (5), 359−363.

(60) Häser, M.; Ahlrichs, R. Improvements on the direct SCF method. *J. Comput. Chem.* **1989**, *10* (1), 104−111.
(61) Hättig, C. Geometry optimizations with the coupled-cluster model CC2 using the resolution-of-the-identity approximation. *J. Chem. Phys.* **2003**, *118* (17), 7751−7761.
(62) Hättig, C.; Hellweg, A.; Köhn, A. Distributed memory parallel implementation of energies and gradients for second-order Møller−Plesset perturbation theory with the resolution-of-the-identity approximation. *Phys. Chem. Chem. Phys.* **2006**, *8* (10), 1159−1169.
(63) Hättig, C.; Weigend, F. CC2 excitation energy calculations on large molecules using the resolution of the identity approximation. *J. Chem. Phys.* **2000**, *113* (13), 5154−5161.
(64) Hoffmann, R. An Extended Hückel Theory. I. Hydrocarbons. *J. Chem. Phys.* **1963**, *39* (6), 1397−1412.
(65) Weigend, F.; Ahlrichs, R. Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy. *Phys. Chem. Chem. Phys.* **2005**, *7* (18), 3297−3305.
(66) Weigend, F.; Häser, M. RI-MP2: First derivatives and global consistency. *Theor. Chem. Acc.* **1997**, *97* (1), 331−340.
(67) Weigend, F.; Häser, M.; Patzelt, H.; Ahlrichs, R. RI-MP2: Optimized auxiliary basis sets and demonstration of efficiency. *Chem. Phys. Lett.* **1998**, *294* (1), 143−152.
(68) *PyTorch*; https://pytorch.org/.
(69) *bwForCluster - BinAC*; https://wiki.bwhpc.de/e/BinAC.
(70) *The PyMOL Molecular Graphics System, Version 3.1*; Schrödinger, LLC, 2015.

https://doi.org/10.1021/acs.jcim.5c03249
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX