bioRxiv preprint doi: https://doi.org/10.1101/2022.01.24.468575; this version posted January 25, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Observing noncovalent interactions in experimental electron density for macromolecular systems: A novel perspective for protein-ligand interaction research

Kang Ding†, Shiqiu Yin†, Zhongwei Li†, Shiju Jiang, Yang Yang, Wenbiao Zhou*, Bo Huang* and Yingsheng Zhang

Affiliation: Beijing StoneWise Technology Co Ltd., Haidian street #15, Haidian district, Beijing, China

†Equal contributors

**Corresponding authors:*

Email: huangbo@stonewise.cn, zhouwenbiao@stonewise.cn

Key words: NCI, Experimental electron density, machine learning

Running title: Experimental NCI database and application

Abstract

The Protein Data Bank (PDB) contains a massive amount of experimental electron density (ED) data. Such data are traditionally used to determine atomic coordinates. We report for the first time the use of experimental ED in the PDB for modeling of noncovalent interactions (NCIs) for protein–ligand complexes. Our methodology is based on the reduced electron density gradient (RDG) theory describing intermolecular NCI by ED and its first derivative. We established a database named Experimental NCI Database (ExptNCI; http://ncidatabase.stonewise.cn/#/nci) containing ED saddle points, indicating ~200,000 NCIs from over 12,000 protein–ligand complexes. The value of such data is demonstrated in a usage case of understanding amide–π interaction geometry in protein-ligand binding system by using the database to facilitate quantum mechanics-based potential energy landscape scan. In summary, the database provides details on experimentally observed NCIs for protein-ligand complexes, and can support future studies on rarely documented NCIs. The potential of fueling artificial intelligence algorithm development by using the database is also discussed.

Introduction

Noncovalent interactions (NCIs) govern protein–ligand interactions and are critical for understanding the determinants affecting ligand-binding affinity. To achieve a deep understanding of NCIs, many protein–ligand interaction databases have been established in the last decade[2-8]. Two types of technologies are mainly applied to build such databases: 1. Structure-based data mining and 2. Quantum mechanical (QM) methods-powered computation. For the first type, protein–ligand complex structures in the Protein Data Bank (PDB) are used as the main source, and different indices, such as distance, angle, exposed surface, and line-of-sight statistics, are used to depict the possibility of NCI between a pair of atoms or two groups[9-11]. For the second type, different levels of QM methods ranging from semiempirical to Coupled-cluster singles-doubles-and-triples wave function (CCSD(T)), are used to quantify the interaction energy of small model complexes[6, 12]. The two technologies together have contributed greatly to the development of rules for the recognition of classical NCIs, such as hydrogen bonds, halogen bonds, salt bridges, and π-π stacking. To further expand the ability to recognize and quantify the entire spectrum of NCIs in highly complicated polarization environments such as protein–ligand binding systems and protein–protein interaction systems, we need to address the gap in direct evidence of NCI between two proximal atoms

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.24.468575; this version posted January 25, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

in macromolecule systems, caused by the limitation of applying quantum mechanics for large systems and by the uncertainty of atom positions in the structures in PDB: e.g., the absence of hydrogen atoms and errors induced during structure building.

A potential solution for this gap can be found in the field of materials research[13] in studies applying the reduced electron density gradient (RDG) theory[14] in analyzing experimental electron density (ED) derived from the X-ray diffraction of small molecular crystals[15, 16]. Stating the RDG theory in simple terms, NCI can be observed by pinpointing the ED saddle point, i.e. (3,-1) critical points, and further quantified by measuring ED deviation from a homogeneous electron distribution by using density and its first derivative ($s = [1/(2(3\pi^2)^{1/3})]|\nabla\rho|/\rho^{4/3}$). Some researchers have even proved that experimental ED can contribute to optimizing functions for density functional theory (DFT), given the fact that experimental ED is inherently time-averaged while DFT ED represents pure ground-state[13].

Inspired by research on small molecule crystals[13, 15, 16], we have developed a potentially path-breaking procedure to extract critical points from experimental ED for protein–ligand complexes deposited in the PDB. We processed over 12,000 protein–ligand complexes and extracted ~200,000 saddle points. These data were subject to noise reduction by varying the ED resolution and then consolidated into a database named ExptNCI (Experimental NCI Database), which is available through the link http://ncidatabase.stonewise.cn/#/nci. In addition to database construction, we also present a case of using such data for empirical NCI mining. ED saddle points indicating amide–π interactions are extracted and used to support the QM interaction energy landscape scan. The QM result is well aligned with the observed points: 85% of the observed points are covered by the region with energy lower than -1.44 kcal/mol (semiempirical level). Besides the attractive interaction of NH/π, which is consistent with previous research[17], we also found a -2.65 kcal/mol interaction (DFT level) between the edge of the aromatic ring and the amide plane when they interact in a perpendicular "edge-on" geometry.

Results

NCI Observed in Experimental ED of the Protein–Ligand complex

X-ray diffraction (XRD) detects the electron distribution of the target molecule and generates an ED map. By searching for the ED saddle points, we can not only recognize classical NCI, such as hydrogen bonds, π stacking, and halogen bonds, but also find relatively rare NCI such as fluorine interacting with sulfur and methyl interacting with pyridine, as shown in Figure 1a-e. Additionally, because experimental ED

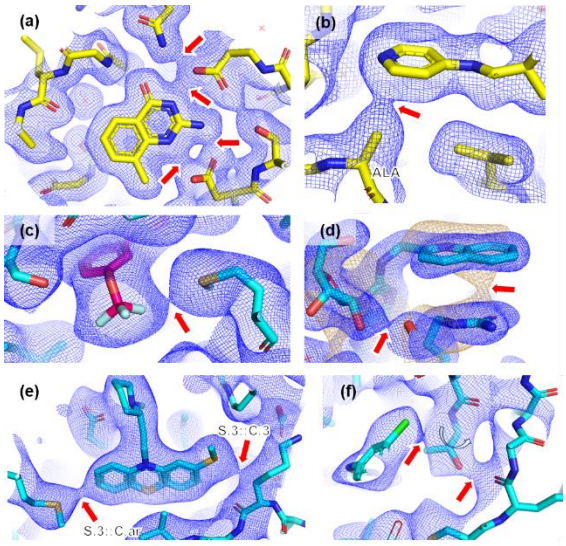


Figure 1. Observing NCI in X-ray diffraction-derived electron density map (2Fo–Fc). Blue mesh indicates 2.5Å resolution. All the maps are sigma scaled and presented at specified counter level. Saddle points are indicated by red arrows. a) Hydrogen bond interaction (PDB: 1S38, map counter level 0.2 sigma); b) Interaction between methyl and aromatic ring (PDB: 1Q8T, map counter level 0.2 sigma); c) interaction between F and methylthio (PDB: 2P4Y, map counter level 0 sigma); d) Weak π stacking revealed in low-resolution electron density map (PDB: 3LDQ; blue mesh indicates 2.5Å map countered at 1.0 sigma; sand yellow mesh indicates 3.5Å resolution map countered at 1.0 sigma); e) Sulfur involved NCI (PDB: 4I1R, map counter level 0.3 sigma); f) Observing NCI under dynamic context caused by the rotation of threonine side chain (PDB:1XKK, map counter level 0 sigma);

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.24.468575; this version posted January 25, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

represents a time-averaged density, some dynamics of the NCI can also be observed (Fig. 1f and Fig. 2).

Another benefit of using XRD ED for NCI detection is that we can emphasize the signals of weak NCI by checking them in low-resolution ED maps generated by only including XRDs at low resolution. The intensity of XRD decreases as the resolution increases, which results in a relatively high signal-to-noise ratio for low-resolution ED maps, which enables us to confirm NCIs by checking them in ED maps at different resolutions (Fig. 2). Doing so not only enables the identification of weak NCIs but also helps to distinguish false-positive NCIs.

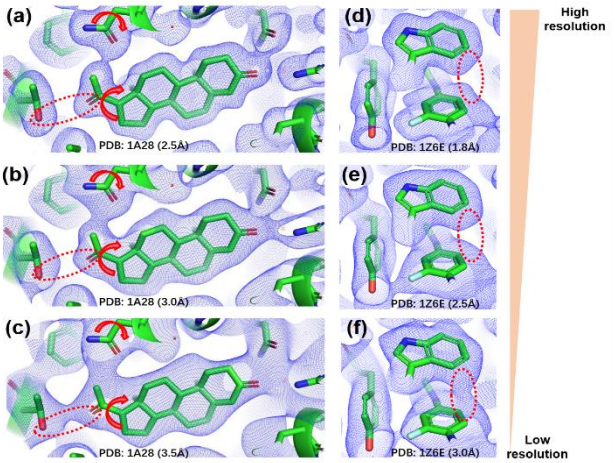


Figure 2. Emphasizing NCI signal in low-resolution ED maps. All the maps are sigma scaled and presented at counter level 0 sigma. Red cycles indicate the relatively weak NCIs which are emphasized in low-resolution ED maps. Hydrogen bonds in a dynamic environment are shown in panels a, b, and c, with 2fo-fc maps for PDB 1A28 at 2.5Å, 3.0Å, and 3.5Å resolution, respectively. Red arrows indicate the rotation of the groups causing the dynamics. π stacking contacts are shown in panels d, e, and f, with 2fo-fc maps for PDB 1Z6E at 1.8Å, 2.5Å, and 3.0Å resolution, respectively.

In addition to using saddle points as a general indicator for recognizing NCI, we also used RDG in experimental ED for a more comprehensive NCI descriptor. Both repulsive and attractive interactions can be identified and visualized, as shown in Figure 3. Specifically, a spike in the RDG vs. sign(λ_2)ρ plot indicates the presence of NCI (Fig. 3b), with the location of the spike on the negative side of the horizontal axis indicating attractive interaction and that on the positive side of the axis indicating repulsive interaction[14].

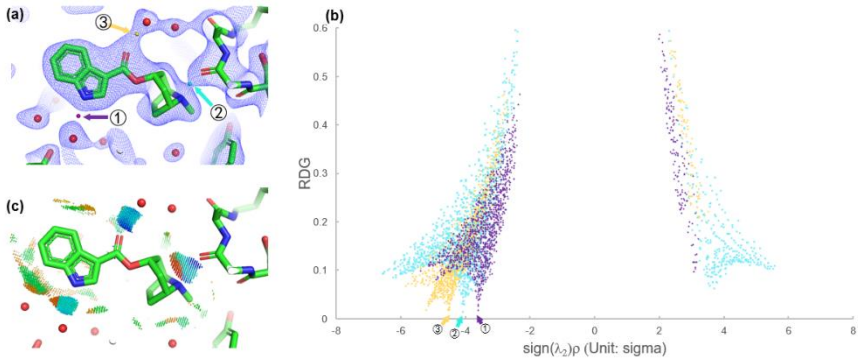


Figure 3. Depicting NCI with RDG in experimental ED for protein–ligand complex (PDB: 2WNC). a) Saddle points detected in 2fo–fc map (counter level 1.0 sigma). Three saddle points indicating three hydrogen bonds are respectively indicated by yellow, purple, and cyan arrows; b) Plots of RDG versus electron density multiplied by the sign of the second Hessian eigenvalue for NCIs indicated in panel (a). Because the ρ here is sigma scaled, to avoid negative value, all the ρ values used for calculating RDG and sign(λ_2)ρ have their value added by 3. Spikes indicating three hydrogen bonds are indicated by arrows. All the dots on the scatter plot are colored according to their positions in real space. In detail, the dots within 1 Å of the saddle points 1, 2, and 3 are colored in yellow, purple, and cyan, respectively. c) RDG-based NCI isosurface showing the ligand–pocket interaction. Regions inside the RDG isosurface at value of 0.2 (arbitrary unit) are indicated with dots, and the dots are colored based on sign(λ_2)ρ using the rainbow scheme, where blue is for large negative values indicating strong attractive interactions and red is for large positive values indicating repulsive interactions.

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.24.468575; this version posted January 25, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

However, there is one limitation of using experimental ED for RDG analysis, which needs to be mentioned. Because of the lack of experimental measures on the forward-scattered reflection swamped by the transmitted beam, which is known as F_{000}, the absolute value of ED is not available for macromolecule crystals. Therefore, the ED maps are contoured on a relative scale, and we had to use a sigma-scaled ρ for the calculation of RDG and $\mathrm{sign}(\lambda_2)\rho$. As a result, the plot in Figure 3b has different scales on the horizontal and vertical axes in arbitrary unit. But the shape of plot, spikes appearing in low-density regions, indicates the occurrence of NCI.

ExptNCI Database Content

The current version of ExptNCI contains a total of 215,397 saddle points extracted from the experimental ED of 12,598 ligand–pocket complex structures in the PDB with resolutions ranging from 2.5 to 4.5, 83% of which have a resolution greater than 2.5 Å. The ED topology information of the saddle points, such as sigma-scaled ρ, RDG, λ_1, λ_2, λ_3, and Laplacian, as well as the structural information of atoms at both ends of the saddle points, such as residual name, element, and its hybridization in the Mol2/Sybyl atom format[18], are included in the database (Table 1). We also included ρ at a low resolution (3.5 Å) at the position of saddle points in a 2.5-Å ED map, for the purpose of using it to distinguish noise from signals of weak NCI. As discussed in the first part of the results section, blurring the map by only including low-resolution data with a relatively high signal-to-noise ratio can emphasize weak NCI. Here, we filtered out false-positive saddle points in a 2.5-Å resolution ED map to check if such points have negative sigma ρ in a 3.5-Å resolution ED map.

After filtering out the saddle points with weak intensity under either 2.5-Å resolution or 3.5-Å resolution, we had 95,532 saddle points left, which accounted for 51% of the originally labeled points (Fig. 4a). Among them, 32% were also recognized as NCI by rules embedded in the widely used software ODDT[19], with hydrogen bonds accounting for the majority (Fig. 4b). For the 68% that were not recognized by ODDT, we made a rough classification based on the properties of the atoms at both ends of the saddle points, as shown in Figure 1c, in which polar interactions (hydrophilic–hydrophilic), aliphatic C…hydrophilic (N/O) interactions, and aromatic …hydrophilic (N/O) interactions accounted for the majority.

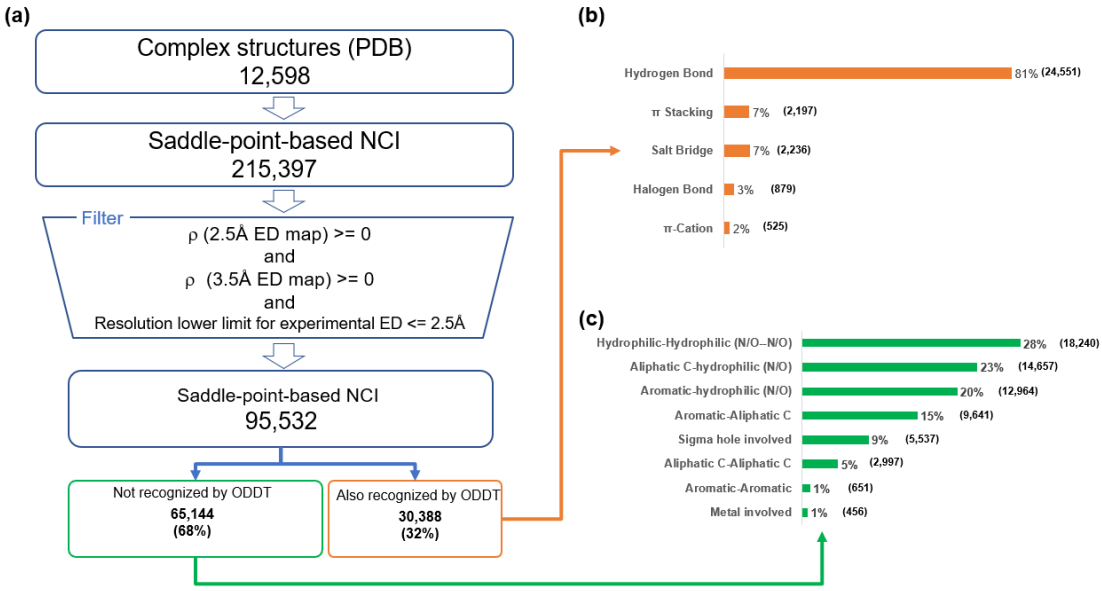


Figure 4. Database construction and dataset profile. a) Database construction workflow; b) Distribution of interaction type for NCIs recognized by both ODDT and ED saddle points; c) Distribution of interaction type for NCIs recognized by ED saddle points but not ODDT

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.24.468575; this version posted January 25, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

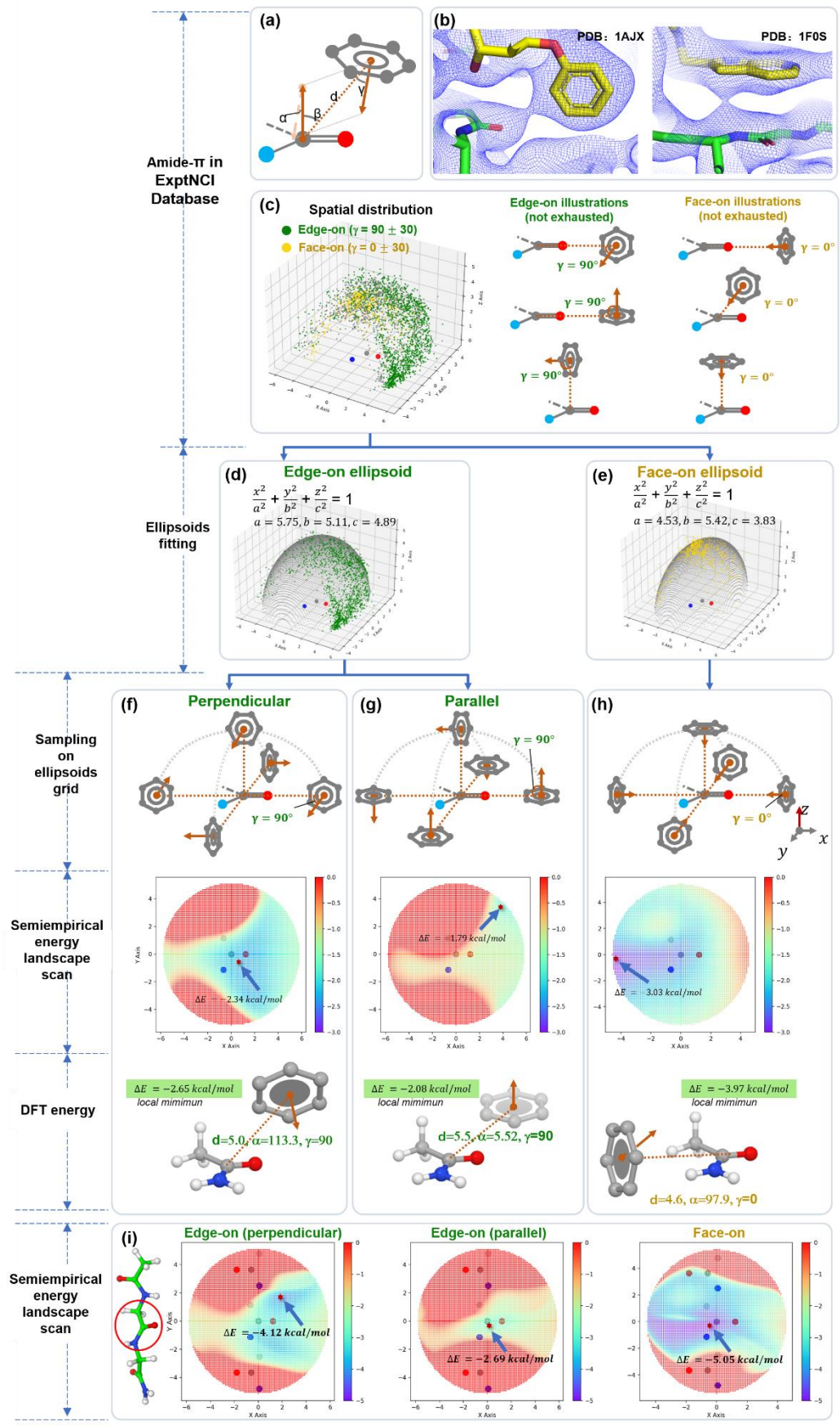
Table 1. List of fields in ExptNCI database

Field	Description
Code	PDB code
LR_id	Ligand Atom Index :: Pocket Atom Index (index start from 0)
L_type	Ligand Atom Type in Mol2/Sybyl type
R_type	Pocket Atom Type in Mol2/Sybyl type
Type	Pocket Atom Index :: Ligand Atom Index in sorted order
LR_type	Pocket Atom Index :: Ligand Atom Index
is_ED_Based	NCI is an ED saddle point-based NCI
Is_ODDT_Rule_Based	NCI is a rule-based NCI recognized by ODDT
ED intensity (2.5A)	ρ, where ρ is sigma scaled ED intensity of the saddle point in 2fo-fc map at 2.5Å
ED intensity modified (2.5A)	ρ +3, where ρ is sigma scaled ED intensity of the saddle point in 2fo-fc map at 2.5Å
ED intensity (3.5A)	ρ, where ρ is sigma scaled ED intensity (in 3.5Å 2fo-fc map) at the position of 2.5Å map saddle point
Distance	Distance between ligand atom and pocket atom
Rule_type	NCI type in rule-based NCI (hbond, salt_bridge, halogen_bond, **π**_stacking, **π**_cation)
is_backbone	NCI occurs on protein backbone
ResNum	Residue number from PDB file
ResAtomName	Residue atom name in PDB file
Resolution	Highest resolution available in PDB
CP_type	(3,-1) indicates saddle point; (3,+1) indicates ring CP
Lambda1 Lambda2 Lambda3	three eigenvalues l_i of the electron-density Hessian (second derivative) matrix, such that ($\lambda_1<=\lambda_2<=\lambda_3$).
Laplacian	$\nabla^2\rho=\lambda_1+\lambda_2+\lambda_3$
RDG	$[1/(2(3\pi^2)^{1/3})]\lvert\nabla\rho\rvert/\rho^{4/3}$
sign(l_2) r	ρ when λ_2 is positive, indicating repulsive interaction; -ρ when λ_2 is negative, indicating attractive interaction
Group	Pocket atom in Protein, Water or Hetatoms
rec_atom_type	Roche atom type (https://doi.org/10.1021/acs.jmedchem.9b01545)

Usage Case：Depicting amide–π interactions in the ligand–protein binding system

Amide–π interactions[17] have been increasingly studied for their involvement in the binding of drug molecules to target proteins[20-22]. Most of the previous studies focused on how the plane of the arene ring interacts with the amide,[17, 21-23] and therefore can be classified as focusing on face-on geometry, a configuration with an γ around 0° in a coordinate system, as shown in figure 5a. To check whether such face-on geometry represents the majority of amide–π interactions in the protein–ligand binding system, we extracted 3,162 amide-π pairs from the ExptNCI database (details of the list provided in supplementary information). The amide–π pairs were extracted based on the fulfillment of the following requirements: 1. It must have ED saddle points between the aromatic carbon and any atom of the amide group; 2. The ED map must have a resolution better than 2.5 Å (examples shown in Fig. 5b). Those with saddle points between C=O and hetero atoms in the aromatic ring were excluded so that classical hydrogen bonds are not included in the analysis. By plotting the spatial distribution of the aromatic ring center relative to the carbon atom of the amide plane and coloring the distribution with a γ-related color scheme, it is interesting to see that most of the interactions had γ values of around 90°, which indicates an edge-on geometry (Fig. 5c). It is also interesting to note that face-on and edge-on interactions occur on two ellipsoids with different radii (Fig. 5d and 5e).

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.24.468575; this version posted January 25, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



bioRxiv preprint doi: https://doi.org/10.1101/2022.01.24.468575; this version posted January 25, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 5. Using experimental ED data to support the profiling of amide–π interaction. a) Examples of amide–π interaction identified by ED saddle points. 2fo–fc map is countered at 0.3 sigma; b) Coordinate system of amide–π interaction; c) Spatial distribution of aromatic ring center relative to the carbon atom of the amide plane. Green and yellow indicate edge-on and face-on geometry, respectively. Illustrations of edge-on and face-on are also provided; (d) and (e) are ellipsoids and parameters obtained by fitting edge-on and face-on positions, respectively, to the general equation of an ellipsoid; For (f), (g), and (h), top part represents the sampling scheme of formamide–benzene conformation on the ellipsoids, with edge-on conformation sampled in two ways: perpendicular and parallel; middle part represents GFN2-xTB level energy landscape, with blue arrow pointing to a red star indicating global minimum; bottom part represents the conformation for global minimum on GFN2-xTB energy landscape and its M06-2x/6-311+G(d,p) energy calculated by GAMESS; i) Energy landscape scan for N-Acetyl Glycyl Glycinamide. The ellipsoid is with respect to the amide group indicated by the red cycle.

To further investigate the interaction geometry for amide–π, we identified the ellipsoids for face-on and edge-on geometry by fitting the aromatic center positions of the two types of geometry to the general equation of an ellipsoid (Fig. 5d and 5e). We then computed the GFN2-xTB level energy landscape based on the fitted ellipsoids by using a formamide–benzene model system (Fig. 5f, 5g, and 5h). For edge-on geometry (i.e., γ=90), the interaction is favored when benzene approaches the amide plane from the top of C=O perpendicularly (Fig. 5f and 5g), with a minimum interaction energy of -2.65 kcal/mol calculated using M06-2x/6-311+G(d,p). For face-on geometry (i.e., γ=0), the result of our energy landscape scan is consistent with previous studies[17], showing a favored interaction of NH/π and a repulsive interaction of C=O/π, as shown in Figure 5h. The same approach was also applied to the amide group in a tripeptide to simulate the situation in the protein (Fig. 5i). The computed energy landscape enjoyed a decent match to the spatial distribution of the observed amide–π interactions extracted from ExptNCI, with 85% of the latter covered by the former region with energy lower than -1.44 kcal/mol.

In summary, the use of observed ED saddle points for NCI description is demonstrated in this case through its support for an energy landscape scan.

Discussion

X-ray diffraction provides an experimental ED map that contains massive amounts of information. Partial information is effectively interpreted into atom coordinates and this information is entered in the PDB. However, in addition to atom coordinates, there is still plenty of information hidden in the experimental ED maps. For the first time, we extracted NCI signals from the ED maps and used it to establish the ExptNCI database.

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.24.468575; this version posted January 25, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

When exploring the ExptNCI database, users should check the three following aspects if some seemingly unusual NCIs are found: 1. Check if the structure is correctly determined, which can be judged by checking whether there are positive or negative densities around the NCI region of interest in the Fo–Fc map; 2. Check if low resolution causes merging of saddle points. An ED map becomes less detailed when the resolution is low, and two proximal saddle points may merge into one in a low-resolution ED map. As shown in Figure 6, just because there is only one saddle point between C=O and C=O in a 2.7-Å resolution ED map, does not necessarily indicate the existence of NCI between the two *sp*^{2} oxygen atoms. In other words, the case in Figure 6 resulted from the merging of two saddle points standing for two individual classical hydrogen bonds; 3. Check if there are any dynamics that can make the interaction more reasonable, e.g., the flip of the side chain for Gln, Asn, and His.

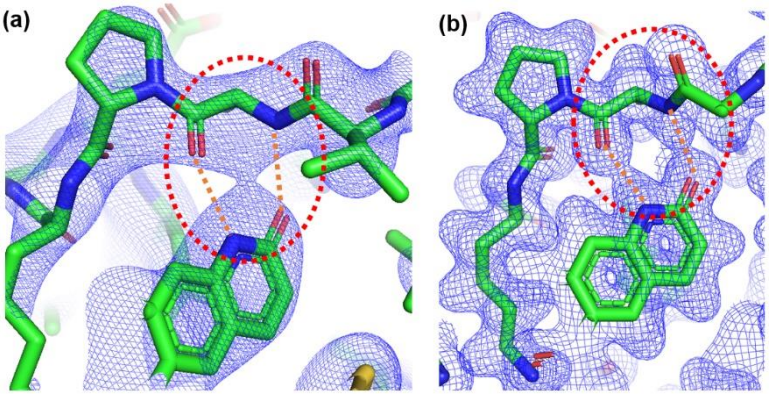


Figure 6. Merging saddle points in low-resolution ED maps (PDB: 6MA1). a) Experimental ED 2Fo–Fc map at 2.7-Å resolution shown at counter level of 1.4 sigma. Two classical hydrogen bonds, indicated by orange dash line, exist in the cycled region, but only one saddle point is observed; b) GFN2-xTB calculated electron density, showing two saddle points at the same region. The map is countered at 0.03 $e^-/Å^3$. The empirical QM calculation is conducted using xtb[1].

How can the data be further improved in terms of quantity and quality? We consider two directions: the first is to expand the scale of the database by extracting NCIs from the interface of protein–protein interactions (PPI). This may allow us to achieve a more detailed understanding of the interaction fingerprint and ultimately benefit peptide/protein design. The second direction is to improve the accuracy of the data by solving multi-crystal variance, which is a problem caused by the lack of absolute ED values for macromolecule crystals. Such a challenge could be tackled by converting the ED values from the sigma-scaled density to the number of electrons. Previous studies that were aimed at measuring the quality of structures in PDB by analyzing ED can serve as a good starting point [24, 25].

Including the experimental ED saddle point intensity as NCI information can also be considered as a solution to support artificial intelligence-based protein–ligand binding prediction. Although experimental NCI is not always available as input—because most often, pocket–ligand complexes are generated by docking or molecular dynamics and, thus, lack experimental ED—we can build two machine learning models to first predict NCIs from a given protein–ligand complex structure, and then use the predicted NCIs to facilitate ligand binding affinity prediction.

In addition to providing more data resources, describing NCI from the perspective of crystallography ED also inspired us to consider leveraging crystallography as a solution for molecular representation for machine learning models. To date, the majority of attempts by researchers to find molecular representations have been in real space, and many reports have been made using strings, molecular graphs, molecular matrices, potential fields, and atom density fields[26]. However, an ideal representation comprehensively reflecting physical and chemical information, friendly to mathematics, and supported with plenty of experimental data available for AI model training is still absent. By applying crystallography

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.24.468575; this version posted January 25, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

theory, we can further expand the attempt in reciprocal space (i.e., frequency domain) and take a big step forward to the realization of the ideal representation for molecules. To be more specific, we apply Fourier transformation (FT) on the atomic coordinates to transfer the information from real space to the frequency domain, and then we apply reverse FT on the frequency domain to bring back the information to real space as ED. By varying the resolution when conducting reverse FT in the frequency domain, we can obtain ED in real space with different levels of detail, emphasizing scaffold, atom, or even bond properties. Unlike graphs composed of vertices and edges, such representations fill the space in a continuously differentiable manner, which is favored by the CNN model. Unlike other 3D molecular representations, such representations are naturally associated with a large amount of testing data: the experimental ED deposited in the PDB. We have already tested such molecular representation on a 3D molecule generation model and have seen some promising results that will be reported later.

In summary, there is a massive amount of information in the experimental ED maps deposited in the PDB. The usage of only part of that information has created our current understanding of protein structures. We hope that our work can shed some light on leveraging experimental ED maps to further understand NCI in the macromolecular system and on combining crystallography and AI from the perspective of providing reliable data sources and exploring better representation of molecules.

Methods

Database construction

1. Experimental ED map processing and critical point labeling

All coordinates and map coefficients were obtained from PDB-REDO[27]. ED maps covering ligands and pocket residues within 5 Å of the ligands were synthesized at multiple resolutions using Phenix[28]. The maps were stored in the xplor format with a 0.15-Å grid interval. The critical points were labeled using the following procedure:

1) Ligand/receptor atom pairs with a distance less than 5 Å were identified and the midpoint was set as the origin
2) The RDG value of all the grids was calculated within 1 Å of the origin
3) The gird point with local minimum RDG was found and marked as a saddle point candidate
4) For all the saddle point candidates, the eigenvalue of the Hessian matrix was calculated and sorted such that $\lambda_3 > \lambda_2 > \lambda_1$. If the eigenvalues did not fulfill the criteria of $\lambda_3 > 0 > \lambda_2 > \lambda_1$, the candidate was discarded
5) If there were two saddle point candidates less than 0.5 Å from each other, the one with relatively weaker intensity was discarded.

2. Atom property annotation

The topology of ligands from PDB entries was curated by RDKit with isosteric SMILES from RCSB Ligand-Expo, and other ligands with missing data were curated using OpenBabel. The Mol2/Sybyl atom types of pockets and ligands in the database were annotated using OpenBabel and PyBel packages, and the rule-based molecular interactions in the database were analyzed and classified using the ODDT software package (version 0.7).

3. Web interface implementation

The database website was developed with a Java backend. The ligand similarity search or substructure search in the database was developed using RDKit, and NCI information was stored and queried through MySQL. NGL.js was implemented to display the receptor–ligand complex and the ED map.

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.24.468575; this version posted January 25, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Amide–π interaction model

To avoid including O..N.ar hydrogen bonds, only amide–π systems with ED saddle points between C/N/O on the protein backbones and C.ar on ligands were subject to our analysis. To profile the spatial distribution of aromatic ring centers, all the amide groups of interest were superimposed and placed on the X-Y plane with a uniform orientation (Fig. 5b), and all the aromatic centers of the amide–π systems were plotted in the Z-positive sector, given that the amide plane is a mirror plane.

Four parameters including angles α, β, γ, and distance d are defined as shown in Figure 5b to describe amide–π geometry, where the angle α is used to describe whether the π system is parallel (α=0 °±30° or 180 °±30°) or perpendicular (α=90±30°) to the amide plane, the angle γ is used to describe whether the aromatic ring center is facing toward the amide group in a "face-on" geometry (γ=0±30°), or showing its edge toward the amide group in an "edge-on" geometry (γ=90±30°).

Ellipsoids for face-on and edge-on geometry were identified by fitting the aromatic center position for the two types of geometry to the general equation of an ellipsoid (Fig. 5d, 5e). Then the fitted ellipsoids are represented by grids with an interval of 0.1 Å along both X and Y axes. To scan the interaction energy landscape based on fitted ellipsoids for benzene and formamide systems, we first determined the zero-point energy (-29.70 kcal/mol) by applying GFN2-xTB calculation on a benzene–formamide complex with distance of 50 Å between the two groups. Then, we placed benzene on the face-on grid in the pose where γ equals 0° to obtain a face-on geometry complex subset (Fig. 5h). For the edge-on geometry complex subset, when we placed a benzene group on the grid of edge-on ellipsoids, there were two types of poses that fulfilled the requirement of γ=90°. Therefore, we divided the edge-on geometry into perpendicular-edge-on and parallel-edge-on subtypes. In detail, we used a plane defined by the Z axis and the vector connecting the amide carbon to the benzene center to distinguish the two subtypes: if the norm of benzene was in the above-defined plane, then it was a parallel-edge-on sub-type (Fig. 5f); if the norm of benzene was perpendicular to the above-defined plane, then it was a perpendicular-edge-on sub-type (Fig. 5g). The energy landscapes for the geometries of face-on, parallel-edge-on, and perpendicular-edge-on were synthesized by calculating GFN2-xTB energy and then subtracting the zero-point energy from it for the complexes on the corresponding grids. Complexes representing the global minimum of the three energy landscapes were also subject to the M06-2x/6-311+G(d,p) calculation using GAMESS to obtain the DFT level energy. The energy landscape scan for benzene interacting with amide groups in the context of tripeptide was conducted in a similar way using a GFN2-xTB-optimized N-acetyl glycyl glycinamide as a starting point. Because the optimized molecule is not subject to mirror symmetry, we scanned the entire ellipsoid and combined the upper and lower halves by overlapping the grids of the two parts and using the lower energy on the two overlapped grids as the final value to compose the energy landscape.

Software for figures and tables

The structure and ED figures were made using Pymol[29]. Statistical analysis was performed using Pandas[30] and Numpy packages[31]. Scatter plots were constructed using Matplotlib[32] and Inkscape[33].

Data availability

http://ncidatabase.stonewise.cn/#/nci

Code availability

Available upon request.

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.24.468575; this version posted January 25, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Author contributions

B. H. conceived the idea. Y. Z. provided instructions for all experiments. W. Z. provided instructions on AI models. K. D. constructed the database and built the NCIScore model. S. Y. developed the saddle point labeling script and constructed the 3DCNN model for saddle point prediction. Z. L. supported the saddle point labeling and performed quantum mechanics calculations. S. J. implemented the web interface for the database. Y. Y. designed the web interface.

Competing interests

The authors declare no competing interests.

Acknowledgment

This work was supported by StoneWise. This work was also partially supported by the Beijing Municipal Science & Technology Commission project Z211100003521001.

References

1. Bannwarth, C.; Ehlert, S.; Grimme, S., GFN2-xTB-An Accurate and Broadly Parametrized Self-Consistent Tight-Binding Quantum Chemical Method with Multipole Electrostatics and Density-Dependent Dispersion Contributions. *J Chem Theory Comput* **2019**, 15, 1652-1671.

2. Anand, P.; Nagarajan, D.; Mukherjee, S.; Chandra, N., PLIC: protein-ligand interaction clusters. *Database (Oxford)* **2014**, 2014, bau029.

3. Angles, R.; Arenas-Salinas, M.; García, R.; Reyes-Suarez, J. A.; Pohl, E., GSP4PDB: a web tool to visualize, search and explore protein-ligand structural patterns. *BMC Bioinformatics* **2020**, 21, 85.

4. Gallina, A. M.; Bisignano, P.; Bergamino, M.; Bordo, D., PLI: a web-based tool for the comparison of protein-ligand interactions observed on PDB structures. *Bioinformatics* **2012**, 29, 395-397.

5. Inhester, T.; Rarey, M., Protein–ligand interaction databases: advanced tools to mine activity data and interactions on a structural level. **2014**, 4, 562-575.

6. Jurecka, P.; Sponer, J.; Cerný, J.; Hobza, P., Benchmark database of accurate (MP2 and CCSD(T) complete basis set limit) interaction energies of small model complexes, DNA base pairs, and amino acid pairs. *Physical chemistry chemical physics : PCCP* **2006**, 8, 1985-93.

7. Murakami, Y.; Omori, S.; Kinoshita, K., NLDB: a database for 3D protein-ligand interactions in enzymatic reactions. *Journal of structural and functional genomics* **2016**, 17, 101-110.

8. Rezac, J., Non-Covalent Interactions Atlas Benchmark Data Sets: Hydrogen Bonding. *J Chem Theory Comput* **2020**, 16, 2355-2368.

9. Ferreira de Freitas, R.; Schapira, M., A systematic analysis of atomic protein-ligand interactions in the PDB. *Medchemcomm* **2017**, 8, 1970-1981.

10. Kuhn, B.; Gilberg, E.; Taylor, R.; Cole, J.; Korb, O., How Significant Are Unusual Protein-Ligand Interactions? Insights from Database Mining. *J Med Chem* **2019**, 62, 10441-10455.

11. Xu, Z.; Zhang, Q.; Shi, J.; Zhu, W., Underestimated Noncovalent Interactions in Protein Data Bank. *J Chem Inf Model* **2019**, 59, 3389-3399.

12. Hobza, P., Calculations on noncovalent interactions and databases of benchmark interaction energies. *Acc Chem Res* **2012**, 45, 663-72.

13. Kasai, H.; Tolborg, K.; Sist, M.; Zhang, J.; Hathwar, V. R.; Filso, M. O.; Cenedese, S.; Sugimoto, K.; Overgaard, J.; Nishibori, E.; Iversen, B. B., X-ray electron density investigation of chemical bonding in van der Waals materials. *Nat Mater* **2018**, 17, 249-252.

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.24.468575; this version posted January 25, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

14. Johnson, E. R.; Keinan, S.; Mori-Sanchez, P.; Contreras-Garcia, J.; Cohen, A. J.; Yang, W., Revealing noncovalent interactions. *J Am Chem Soc* **2010**, 132, 6498-506.

15. Saleh, G.; Gatti, C.; Lo Presti, L., Non-covalent interaction via the reduced density gradient: Independent atom model vs experimental multipolar electron densities. *Computational and Theoretical Chemistry* **2012**, 998, 148–163.

16. Saleh, G.; Gatti, C.; Lo Presti, L.; Contreras-Garcia, J., Revealing non-covalent interactions in molecular crystals through their experimental electron densities. *Chemistry* **2012**, 18, 15523-36.

17. Imai, Y. N.; Inoue, Y.; Nakanishi, I.; Kitaura, K., Amide-pi interactions between formamide and benzene. *J Comput Chem* **2009**, 30, 2267-76.

18. Clark, M.; Cramer III, R. D.; Van Opdenbosch, N., Validation of the general purpose tripos 5.2 force field. **1989**, 10, 982-1012.

19. Wojcikowski, M.; Zielenkiewicz, P.; Siedlecki, P., Open Drug Discovery Toolkit (ODDT): a new open-source player in the drug discovery field. *J Cheminform* **2015**, 7, 26.

20. Krone, M. W.; Travis, C. R.; Lee, G. Y.; Eckvahl, H. J.; Houk, K. N.; Waters, M. L., More Than pi-pi-pi Stacking: Contribution of Amide-pi and CH-pi Interactions to Crotonyllysine Binding by the AF9 YEATS Domain. *J Am Chem Soc* **2020**, 142, 17048-17056.

21. DeFrees, K.; Kemp, M. T.; ElHilali-Pollard, X.; Zhang, X.; Mohamed, A.; Chen, Y.; Renslo, A. R., An Empirical Study of Amide-Heteroarene pi-Stacking Interactions Using Reversible Inhibitors of a Bacterial Serine Hydrolase. *Org Chem Front* **2019**, 6, 1749-1756.

22. Bootsma, A. N.; Wheeler, S. E., Stacking Interactions of Heterocyclic Drug Fragments with Protein Amide Backbones. *ChemMedChem* **2018**, 13, 835-841.

23. Harder, M.; Kuhn, B.; Diederich, F., Efficient stacking on protein amide fragments. *ChemMedChem* **2013**, 8, 397-404.

24. Lang, P. T.; Holton, J. M.; Fraser, J. S.; Alber, T., Protein structural ensembles are revealed by redefining X-ray electron density noise. *Proc Natl Acad Sci U S A* **2014**, 111, 237-42.

25. Yao, S.; Moseley, H. N. B., A chemical interpretation of protein electron density maps in the worldwide protein data bank. *PLoS One* **2020**, 15, e0236894.

26. Musil, F.; Grisafi, A.; Bartok, A. P.; Ortner, C.; Csanyi, G.; Ceriotti, M., Physics-Inspired Structural Representations for Molecules and Materials. *Chem Rev* **2021**, 121, 9759-9815.

27. Joosten, R. P.; Long, F.; Murshudov, G. N.; Perrakis, A., The PDB_REDO server for macromolecular structure model optimization. *IUCrJ* **2014**, 1, 213-20.

28. 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.; Moriarty, N. W.; Oeffner, R. D.; Poon, B. K.; Prisant, M. G.; Read, R. J.; Richardson, J. S.; Richardson, D. C.; Sammito, M. D.; Sobolev, O. V.; Stockwell, D. H.; Terwilliger, T. C.; Urzhumtsev, A. G.; Videau, L. L.; Williams, C. J.; Adams, P. D., Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix. *Acta Crystallogr D Struct Biol* **2019**, 75, 861-877.

29. Schrodinger, LLC, In; 2015.

30. McKinney, W., Data Structures for Statistical Computing in Python. *Proceedings of the 9th Python in Science Conference* **2010**, 51-56.

31. Harris, C. R.; Millman, K. J.; van der Walt, S. J.; Gommers, R.; Virtanen, P.; Cournapeau, D.; Wieser, E.; Taylor, J.; Berg, S.; Smith, N. J.; Kern, R.; Picus, M.; Hoyer, S.; van Kerkwijk, M. H.; Brett, M.; Haldane, A.; del Río, J. F.; Wiebe, M.; Peterson, P.; Gérard-Marchant, P.; Sheppard, K.; Reddy, T.; Weckesser, W.; Abbasi, H.; Gohlke, C.; Oliphant, T. E., Array programming with NumPy. *Nature* **2020**, 585, 357-362.

32. Hunter, J. D., Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering* **2007**, 9, 90-95.

33. InkscapeProject, Inkscape 0.92.5. **2020**.