PERSPECTIVE

ELSEVIER

feature

Toward structure-based drug design against the epidermal growth factor receptor (EGFR)

Yazan Haddad[1,2], Marek Remes[1,2], Vojtech Adam[1,2] and Zbynek Heger[1,2], heger@mendelu.cz

Most of the available crystal structures of epidermal growth factor receptor (EGFR) kinase domain, bound to drug inhibitors, originated from ligand-based drug design studies. Here, we used variations in 110 crystal structures to assemble eight distinct families highlighting the C-helix orientation in the N-lobe of the EGFR kinase domain. The families shared similar mutational profiles and similarity in the ligand R-groups (chemical composition, geometry, and charge) facing the C-helix, mutation sites, and DFG domain. For structure-based drug design, we recommend a systematic decision-making process for choice of template, guided by appropriate pairwise fitting and clustering before the molecular docking step. Alternatively, the binding site shape/volume can be used to filter and select the compound libraries.

Drug design (ligand-based versus structure-based)

Over the past few decades, drug design has shifted from a computer-aided (i.e., computer graphics [1]) paradigm into a computer-directed or guided process. Depending on the elementary molecule used for computational design, two state-of-the-art strategies have advanced in this field, namely ligand-based and structure-based design. On the ligand-based side, quantitative structure–activity relationship (QSAR) was successfully used for nearly half a century to build reliable statistical models to predict the properties of new chemical compounds. The method is based on the study of quantitative relationships between physicochemical properties of chemical compounds and their biological activities [2]. Structure-based drug design is concerned with how

ligands bind to the protein structure and methods of binding energy estimation. The historic analogy of protein and ligand with 'lock and key' was later replaced with the 'hand and glove' concept to mimic the flexibility of both the protein and ligand [3]. However, covalent inhibitors (in which the glove is permanently linked to the hand) need to first form a stable noncovalent complex in their search for optimal conditions to form covalent links in the final complex [4]. A structure-based virtual screening is performed via bottleneck rounds of high-throughput characterization, analysis, filtering, and selection. The protocol is summarized as follows: (i) selection of target protein structure(s) from crystal structure repositories (e.g., Protein Data Bank; PDB) or from computational models constructed via homology modeling or molecular dynamics

(MD) simulations; (ii) characterization of the binding site; (iii) compound library construction, which involves several processes of characterization, filtering, and clustering; (iv) molecular docking of target with compounds supplemented with known actives and decoys followed by scoring; and (v) final evaluation and validation [5]. Given the dynamic nature of targets (which is required for them to function in physiological conditions), the binding site of the target–ligand complex is often adapted to new conformations involving either the protein side-chains or even the backbone [6].

Many specialized methods for optimizing and improving the accuracy of ligand-based and structure-based drug design have been developed, such as scaffold hopping and R-group replacement of lead compounds [7], the introduction of machine learning to QSAR for

Features • PERSPECTIVE

Please cite this article in press as:

modeling of chemical features in compound libraries [8], and rotamer dynamics for monitoring side-chain flexibility in the molecular simulations of proteins [9]. The most accurate predictions can arise from using a MD simulation approach to screen libraries of compounds, such as the recent attempt to screen 8000 compounds utilizing the SUMMIT supercomputer at Oak Ridge National Laboratory, against the viral spike protein in Coronavirus 2019 (COVID-19) [10]. However, such an extreme amount of computational resource is not available to the average researcher. Docking software has

recently improved in terms of its ability to accommodate side-chain flexibility [3]. Nevertheless, here we show that there are greater changes to be anticipated in the structure-based drug design targeting EGFR, a well-established receptor in many types of cancer.

EGFR tyrosine kinase inhibitors

EGFR (Fig. 1a) is the main druggable oncogenic target in nonsmall cell lung cancer (NSCLC), with nearly 50% of Asian patients and 15% of Caucasian patients presenting with growth-activating mutations. Most of these mutations are

exon 19 deletions and L858R point mutations [11,12]. Almost 10 years have passed since tyrosine kinase inhibitors (TKIs) won the race over 'chemotherapy' in better outcomes of treatment for metastatic NSCLC. Nevertheless, there have been many obstacles, including the development of cellular chemoresistance to TKI therapy. The first-generation of TKIs [gefitinib (ZD1839; Iressa®; AstraZeneca, UK and Teva Pharmaceutical Industries, Israel) and erlotinib (Tarceva® by Astellas Pharma US, Inc., and Genentech, Inc., USA)], which were characterized by reversible binding to EGFR, were
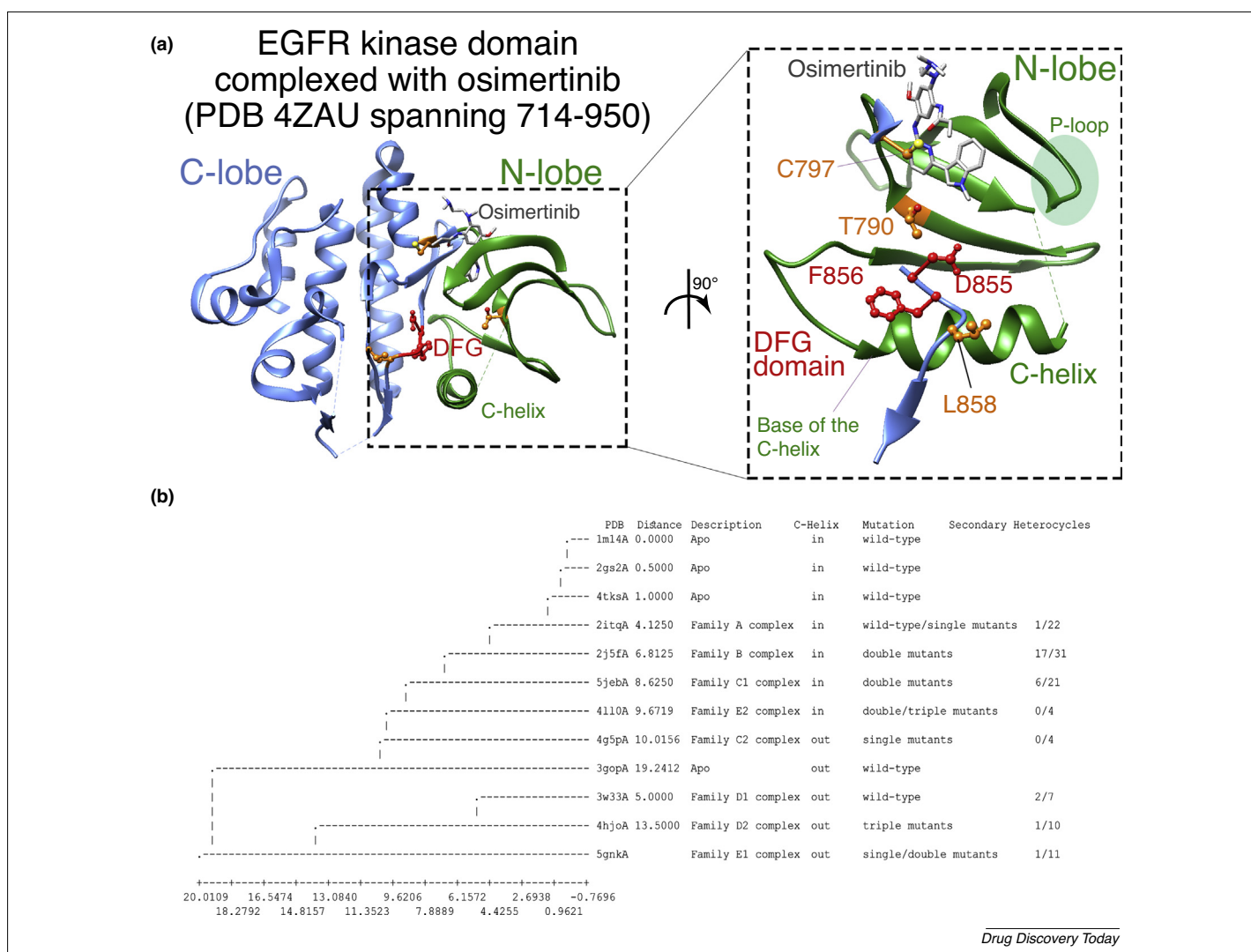
FIGURE 1

N-lobe structural variations in the epidermal growth factor receptor (EGFR) kinase domain. **(a)** 3D structure of the EGFR kinase domain complexed with third-generation drug osimertinib showing the C-lobe (blue) and N-lobe (green). Distinct mutation sites are highlighted in orange, whereas the DFG domain is highlighted in red. **(b)** Dendrogram showing weighted pair-group average clustering of the N-lobe of EGFR based on distance matrix. Structure–structure alignment was used to compare four wild-type apo EGFR with representative structures from classified families of co-complexed EGFR with ligands. Method: The structure–structure multiple-alignment using aln.malign3d() function in Modeller (9.17, r10881) was used to perform throughout (N−1) dynamic programming where each new structure is compared to the average of the previous structures in each new cycle. Default settings were used [off diagonal = 100, gap penalties 3d = (0.0, 1.75), fit atoms = 'CA']. Dendrogram was built in Modeller with a cluster cutoff of −1.0.

Please cite this article in press as:

followed by a second-generation TKIs [afatinib (BIBW2992; Gilotrif® by Boehringer Ingelheim, Germany)] with irreversible binding to EGFR (and HER2 homolog) by TKI binding to Cys797. Yet, resistance to therapy continued. A selective mutation causing a key substitution (T790M) in the kinase domain of EGFR contributed to nearly half of chemoresistant cases. This led to the development of the third-generation TKIs [osimertinib (AZD9291; Tagrisso® by AstraZeneca, UK); rociletinib (CO-1686); olmutinib (HM61713); and nazartinib (EGF816)], which are still challenged by the development of chemoresistance [11,13]. Approximately 10–26% of cases develop resistance to second-line osimertinib treatment because of the EGFR C797S mutation, making it the most common tertiary EGFR mutation. When osimertinib was used as a front-line treatment, C797S mutation was reported in 7% of cases, making it the second most frequent mechanism of drug resistance behind *MET* amplification. Furthermore, C797S mutation is predicted to interfere with binding of other third-generation TKIs, such as rociletinib, olmutinib, and nazartinib [12]. One recent alternative strategy was to develop C797-targeting covalent inhibitors that do not lose their potency (and affinity) in the presence of C797S mutants [14]; however, there was still a need to develop therapeutics using alternative mechanisms of action.

Fourth-generation TKIs were developed by targeting EGFR away from the ATP-binding site. The first allosteric inhibitor compound (called EAI001 or EGFR allosteric inhibitor-1) was discovered by screening a library of ~2.5 million compounds against a mutant peptide derived from EGFR in the presence of ATP. Following optimization for selectivity, the new EAI045 inhibitor was found to have high potency and selectivity as an allosteric, non-ATP competitive inhibitor of L858R/T790M mutant EGFR [15]. The allosteric binding pocket is far from the C797S site and, although EAI045 is ineffective alone because of receptor dimerization, it is promising for use in combination with cetuximab against T790M and C797S mutants [16]. JBJ-04-125-02 is another example of fourth-generation allosteric TKIs that shows promising potency when used in combination with osimertinib [17]; however, when used alone, its efficiency drops because of EGFR dimerization. C797S mutation is not the only reason for TKI drug resistance [12]. Resistance also arises from other EGFR mutations (e.g., G796X), and other acquired alterations, such as gene amplification (e.g., *MET*, *HER2*, etc.), oncogenic fusions, and MAPK-PI3K mutations.

Unfortunately, there is no standard or systematic treatment following the failure of the four generations of TKIs. The navigation of this 'no man's land' is sometimes managed by lytic-inducing strategies or onco-immunologic methods; however, many alternative strategies are under investigation [18]. Recent clinical trials on the combination of chemotherapy and osimertinib have shown promising progress, possibly because of the better toxicological profile of osimertinib compared with previous generation TKIs [19]. A recent meta-analysis on the outcomes of ~4465 patients with EGFR-mutant NSCLC receiving TKI treatment showed that particular cohorts can have more benefits than others; for example, females over males, non-smokers over smokers, and exon 19 deletion over L858R mutation [11]. Whether it is a sign of TKI development slow down or a natural development, 'precision medicine' has become an important step for management of the disease by stratification of patient cohorts into optimized treatment regimens according to molecular, pathophysiological and 'omics profiling [20]. In addition to conventional therapies targeting EGFR, such as TKIs (provisionally via ligand-based drug design) and monoclonal antibodies (e.g., cetuximab, panitumumab, and necitumumab); a third strategy is now targeting EGF directly via EGFR-derived peptide-based inhibitors, anti-EGF vaccines, and single-domain antibodies (nanobodies) [21].

Structure-based drug design of EGFR TKIs has been attracting global attention among researchers (Table S1 in the supplemental information online). NMR and X-ray crystals were used to validate structures of ligands only rather than complexes. Whereas most studies used a single structure (e.g., 1M17) for screening, few studies used two models (wild-type versus mutant) for drug design, such as [22,23]. Interestingly, Sun *et al.* [24] used clusters of 19 EGFR-TKIs complexed structures for docking in an effort to create an ensemble that can be applied in ligand-based drug design. In 40 surveyed studies (Table S1 in the supplemental information online), structure-based drug design was mainly used as a supporting method to estimate ligand pose or interaction energy. Therefore, all MD simulations validations were too short to predict large movements in the EGFR protein.

Comparative overview of EGFR crystal structures

The development of EGFR TKIs is evolving quickly. At the time of writing, >71 900 ISI-impacted papers had been published on EGFR and over 200 X-ray crystalized protein 3D

structures had been studied, mainly complexed with TKI ligands. Here, we analyzed a representative number of EGFR crystal structures to address the following questions: (i) how many different structural variations occur in the EGFR kinase-binding site? (ii) What are their causes based on TKIs and mutations? (iii) How big a role does the scaffold (i.e., core) of the TKIs have in binding compared with the R-groups? (iv) How can inhibitors be classified that reflects their effect? Nearly all the ~40 publications of the 110 crystal structures described here were ligand-based drug design studies complemented with pharmacokinetics and X-ray crystallography to discover the mode of action. Nevertheless, it might be difficult to select which 3D structure to use. In other words, is there a subset of structures that are suitable for each TKI study? Despite being state-of-the-art, X-ray crystallography experiments take a relatively long time and are not expected to cover the most recent trends in EGFR TKIs. Therefore, it is likely that we have only mapped a small fraction of the conformational landscape of known EGFR TKIs.

There are nearly 110 representative EGFR kinase 3D structures spanning the amino acid range 714–950 (numeration according to Uniprot ID P00533-1) and complexed with ligands. The kinase domain is divided into the N-lobe (714–795) and C-lobe (796–950), which clench the TKI ligands (in almost planar geometry) from both sides like a sandwich (Fig. 1a). There were 83 ligands salvaged in these structures, classified according to the core heterocycle of the scaffold into the following classes (numbers in brackets describe the represented PDB structures): one antibiotic (2), two benzimidazoles (3), six furo-pyrimidines (6), two indolocarbazoles (5), seven purines (11), one pyrazine (1), seven pyrazolo-pyrimidines (7), three pyridones (4), 34 pyrimidines (37), one pyrimidopyridone (1), eight pyrrolopyrimidines (13), seven quinazolines (16), one quinoline (2), and two thiazoles (2) (Table S2 in the supplemental information online). Pyrimidine compounds were further classified into the following groups based on their secondary heterocycles: 13 imidazopyridines, one imidazothiazole, two indazoles, two indoles, two pyrazoles, one pyrazolopyridine, two pyridones, four pyrrolopyridines, and seven without secondary heterocycles. Other compounds did not have classifiable heterocycles. This impressive diversity has been achieved in one decade when previously all kinase inhibitors were dominated by quinazoline-based TKIs [25]. Four apo-EGFR entries (resolution in Å) were without mutations or additional compounds: 1M14 (2.6 Å), 2GS2
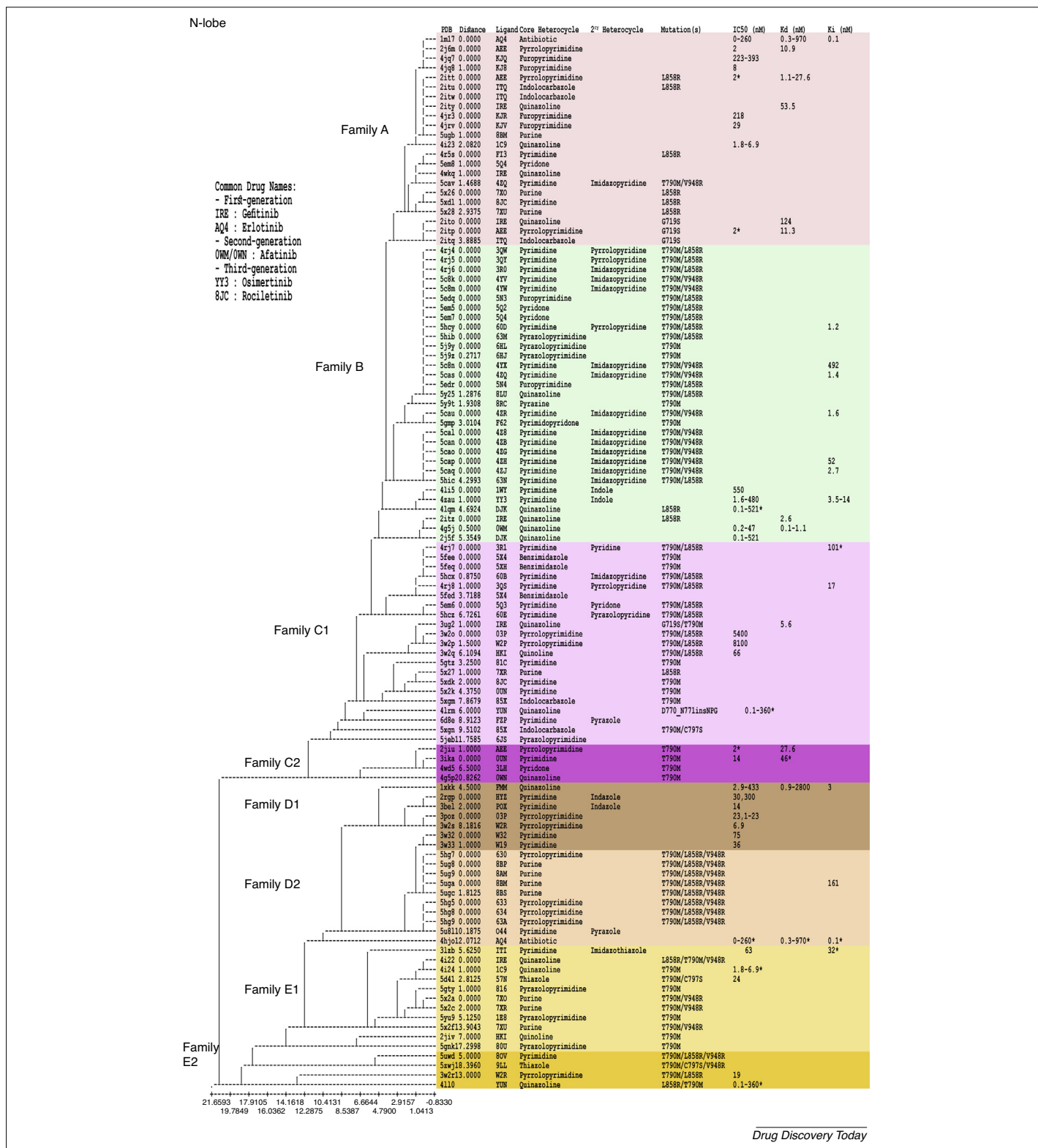
Features • PERSPECTIVE

N-lobe

Common Drug Names:
- First-generation
  IRE : Gefitinib
  AQ4 : Erlotinib
- Second-generation
  0WM/0WN : Afatinib
- Third-generation
  YY3 : Osimertinib
  8JC : Rociletinib

Family	PDB	Distance	Ligand	Core Heterocycle	2ry Heterocycle	Mutation(s)	IC50 (nM)	Kd (nM)	Ki (nM)
A	1m17	0.0000	AQ4	Antibiotic			0-260	0.3-970	0.1
A	2j6m	0.0000	AEE	Pyrrolopyrimidine			2	10.9	
A	4jq7	0.0000	KJQ	Furopyrimidine			223-393		
A	4jq8	1.0000	KJ8	Furopyrimidine			8		
A	2itt	0.0000	AEE	Pyrrolopyrimidine		L858R	2*	1.1-27.6	
A	2itu	0.0000	ITQ	Indolocarbazole		L858R			
A	2itw	0.0000	ITQ	Indolocarbazole					
A	2ity	0.0000	IRE	Quinazoline				53.5	
A	4jr3	0.0000	KJR	Furopyrimidine			218		
A	4jrv	0.0000	KJV	Furopyrimidine			29		
A	5ugb	1.0000	8BM	Purine					
A	4i23	2.0820	1C9	Quinazoline			1.8-6.9		
A	4r5s	0.0000	FI3	Pyrimidine		L858R			
A	5em8	1.0000	5Q4	Pyridone					
A	4wkq	1.0000	IRE	Quinazoline					
A	5cav	1.4688	4ZQ	Pyrimidine	Imidazopyridine	T790M/V948R			
A	5x26	0.0000	7XO	Purine		L858R			
A	5xd1	1.0000	8JC	Pyrimidine		L858R			
A	5x28	2.9375	7XU	Purine		L858R			
A	2ito	0.0000	IRE	Quinazoline		G719S	124		
A	2itp	0.0000	AEE	Pyrrolopyrimidine		G719S	2*	11.3	
A	2itq	3.8885	ITQ	Indolocarbazole		G719S			
B	4rj4	0.0000	3QW	Pyrimidine	Pyrrolopyridine	T790M/L858R			
B	4rj5	0.0000	3QY	Pyrimidine	Pyrrolopyridine	T790M/L858R			
B	4rj6	0.0000	3R0	Pyrimidine	Imidazopyridine	T790M/L858R			
B	5c8k	0.0000	4YV	Pyrimidine	Imidazopyridine	T790M/V948R			
B	5c8m	0.0000	4YW	Pyrimidine	Imidazopyridine	T790M/V948R			
B	5edq	0.0000	5N3	Furopyrimidine		T790M/L858R			
B	5em5	0.0000	5Q2	Pyridone		T790M/L858R			
B	5em7	0.0000	5Q4	Pyridone		T790M/L858R			
B	5hcy	0.0000	60D	Pyrimidine	Pyrrolopyridine	T790M/L858R			1.2
B	5hib	0.0000	63M	Pyrazolopyrimidine		T790M/L858R			
B	5j9y	0.0000	6HL	Pyrazolopyrimidine		T790M			
B	5j9z	0.2717	6HJ	Pyrazolopyrimidine		T790M			
B	5c8n	0.0000	4YX	Pyrimidine	Imidazopyridine	T790M/V948R			492
B	5cas	0.0000	4ZQ	Pyrimidine	Imidazopyridine	T790M/V948R			1.4
B	5edr	0.0000	5N4	Furopyrimidine		T790M/L858R			
B	5y25	1.2876	8LU	Quinazoline		T790M/L858R			
B	5y9t	1.9308	8RC	Pyrazine		T790M			
B	5cau	0.0000	4ZR	Pyrimidine	Imidazopyridine	T790M/V948R			1.6
B	5gmp	3.0104	F62	Pyrimidopyridone		T790M			
B	5cal	0.0000	4Z8	Pyrimidine	Imidazopyridine	T790M/V948R			
B	5can	0.0000	4ZB	Pyrimidine	Imidazopyridine	T790M/V948R			
B	5cao	0.0000	4ZG	Pyrimidine	Imidazopyridine	T790M/V948R			
B	5cap	0.0000	4ZH	Pyrimidine	Imidazopyridine	T790M/V948R			52
B	5caq	0.0000	4ZJ	Pyrimidine	Imidazopyridine	T790M/V948R			2.7
B	5hic	4.2993	63N	Pyrimidine	Imidazopyridine	T790M/L858R			
B	4li5	0.0000	1WY	Pyrimidine	Indole		550		
B	4zau	1.0000	YY3	Pyrimidine	Indole		1.6-480		3.5-14
B	4lqm	4.6924	DJK	Quinazoline		L858R	0.1-521*		
B	2itz	0.0000	IRE	Quinazoline		L858R		2.6	
B	4g5j	0.5000	0WM	Quinazoline			0.2-47	0.1-1.1	
B	2j5f	5.3549	DJK	Quinazoline			0.1-521		
C1	4rj7	0.0000	3R1	Pyrimidine	Pyridine	T790M/L858R			101*
C1	5fee	0.0000	5X4	Benzimidazole		T790M			
C1	5feq	0.0000	5XH	Benzimidazole		T790M			
C1	5hcx	0.8750	60B	Pyrimidine	Imidazopyridine	T790M/L858R			
C1	4rj8	1.0000	3QS	Pyrimidine	Pyrrolopyridine	T790M/L858R			17
C1	5fed	3.7188	5X4	Benzimidazole		T790M/L858R			
C1	5em6	0.0000	5Q3	Pyrimidine	Pyridone	T790M/L858R			
C1	5hcz	6.7261	60E	Pyrimidine	Pyrazolopyridine	T790M/L858R			
C1	3ug2	1.0000	IRE	Quinazoline		G719S/T790M		5.6	
C1	3w2o	0.0000	03P	Pyrrolopyrimidine		T790M/L858R	5400		
C1	3w2p	1.5000	W2P	Pyrrolopyrimidine		T790M/L858R	8100		
C1	3w2q	6.1094	HKI	Quinoline		T790M/L858R	66		
C1	5gtz	3.2500	81C	Pyrimidine		T790M			
C1	5x27	1.0000	7XR	Purine		L858R			
C1	5xdk	2.0000	8JC	Pyrimidine		T790M			
C1	5x2k	4.3750	0UN	Pyrimidine		T790M			
C1	5xqm	7.8679	85X	Indolocarbazole		T790M			
C1	4lrm	6.0000	YUN	Quinazoline		D770_N771insNPG	0.1-360*		
C1	6d8e	8.9123	F2P	Pyrimidine	Pyrazole				
C1	5xqn	9.5102	85X	Indolocarbazole		T790M/C797S			
C1	5jeb	11.7585	6JS	Pyrazolopyrimidine					
C2	2jiu	1.0000	AEE	Pyrrolopyrimidine		T790M	2*	27.6	
C2	3ika	0.0000	0UN	Pyrimidine		T790M	14	46*	
C2	4wd5	6.5000	3LH	Pyridone		T790M			
C2	4g5p	20.8262	0WN	Quinazoline		T790M			
D1	1xkk	4.5000	FMM	Quinazoline			2.9-433	0.9-2800	3
D1	2rgp	0.0000	HYZ	Pyrimidine	Indazole		30,300		
D1	3bel	2.0000	POX	Pyrimidine	Indazole		14		
D1	3poz	0.0000	03P	Pyrrolopyrimidine			23,1-23		
D1	3w2s	8.1816	W2R	Pyrrolopyrimidine			6.9		
D1	3w32	0.0000	W32	Pyrimidine			75		
D1	3w33	1.0000	W19	Pyrimidine			36		
D2	5hg7	0.0000	630	Pyrrolopyrimidine		T790M/L858R/V948R			
D2	5ug8	0.0000	8BP	Purine		T790M/L858R/V948R			
D2	5ug9	0.0000	8AM	Purine		T790M/L858R/V948R			
D2	5uga	0.0000	8BM	Purine		T790M/L858R/V948R			161
D2	5ugc	1.8125	8BS	Purine		T790M/L858R/V948R			
D2	5hg5	0.0000	633	Pyrrolopyrimidine		T790M/L858R/V948R			
D2	5hg8	0.0000	634	Pyrrolopyrimidine		T790M/L858R/V948R			
D2	5hg9	0.0000	63A	Pyrrolopyrimidine		T790M/L858R/V948R			
D2	5u81	10.1875	O44	Pyrimidine	Pyrazole	T790M/L858R/V948R			
D2	4hjo	12.0712	AQ4	Antibiotic			0-260*	0.3-970*	0.1*
E1	3lzb	5.6250	ITI	Pyrimidine	Imidazothiazole	L858R/T790M/V948R	63		32*
E1	4i22	0.0000	IRE	Quinazoline		T790M			
E1	4i24	1.0000	1C9	Quinazoline		T790M/C797S	1.8-6.9*		
E1	5d41	2.8125	57N	Thiazole		T790M/C797S	24		
E1	5gty	1.0000	816	Pyrazolopyrimidine		T790M			
E1	5x2a	0.0000	7XO	Purine		T790M/V948R			
E1	5x2c	2.0000	7XR	Purine		T790M/V948R			
E1	5yu9	5.1250	1E8	Pyrazolopyrimidine		T790M			
E1	5x2f	13.9043	7XU	Purine		T790M/V948R			
E2	2jiv	7.0000	HKI	Quinoline		T790M			
E2	5gnk	17.2998	80U	Pyrazolopyrimidine		T790M			
E2	5uwd	5.0000	8OV	Pyrimidine		T790M/L858R/V948R			
E2	5zwj	18.3960	9LL	Thiazole		T790M/C797S/V948R			
E2	3w2r	13.0000	W2R	Pyrrolopyrimidine		T790M/L858R	19		
E2	4l10		YUN	Quinazoline		L858R/T790M	0.1-360*		

```
+----+----+----+----+----+----+----+----+----+----+----+----+
21.6593  17.9105  14.1618  10.4131   6.6644   2.9157  -0.8330
    19.7849  16.0362  12.2875   8.5387   4.7900   1.0413
```

Drug Discovery Today

FIGURE 2

Dendrogram showing weighted pair-group average clustering of the N-lobe of epidermal growth factor receptor (EGFR) based on a distance matrix. Structure–structure alignment was used to compare 110 3D structures and resulted in two distinct clans that can be divided further into two to three families each. The largest clan of 78 structures was highly similar and divided into family A (22 structures with one or no mutations in EGFR), family B (31 structures with mostly double EGFR mutations), and family C (25 structures with one or two EGFR mutations, divided into subfamily 1 with 21 structures and subfamily 2 with four structures). The other clan was more divergent and was divided into family D (17 structures divided into subfamily 1 with seven wild-type EGFR structures and subfamily 2 with ten triple-mutated EGFR structures) and family E (15 highly divergent structures divided into two subfamilies of ten and five structures). Method: Using the keyword 'EGFR', a RCSB protein databank (www.rcsb.org) database search resulted in 260 structures. Structures were ordered according to best resolution (Å), and then entries that did not cover the kinase domain of EGFR were excluded. Thirteen entries were without mutations and also without

Please cite this article in press as:

(2.8 Å), 3GOP (2.8 Å), and 4TKS (3.2017 Å). The entry 3GOP was chosen as the reference structure because of its length coverage and good resolution. However, it is the only apo-EGFR entry with a C-helix-out (inactive kinase) conformation, whereas the remainder were presented in the C-helix-in (active kinase) conformation (Fig. 1B).

The 110 EGFR 3D structures spanning the full range of the kinase domain were divided into five major family clusters at average weighted distances: 21.4069, 20.6099, 16.6741, and 9.4959 (Fig. S1 in the supplemental information online). There was no clear distribution of ligand classes among these families; however, EGFR mutations were relatively distributed among families in triple, double, single or no EGFR mutations. To identify local structural regions of divergence, the 3D structures of the two lobes in EGFR kinase were compared separately. Surprisingly, the C-lobe displayed relatively low divergence compared with the full-range and the N-lobe structures, with a maximum weighted distance of 5.0684 (Fig. S2 in the supplemental information online). This led to the conclusion that major structural variations must reside in the N-lobe, which requires further investigation.

Of the 110 EGFR kinase N-Lobe crystal structures, two distinct clans (78 and 32 structures) were separated at an average weighted distance of 20.83, with the highest reported root-mean-square deviation (RMSD) against reference of 4.242 Å by 3LZB structure and lowest of 1.438 Å by 2JIT structure (Fig. 2; Table S3 in the supplemental information online) with different C-helix orientations (Fig. 3). Nearly 50 structures within the largest clan were at a distance below 5.0, whereas the other clan showed higher divergence. At least eight distinct families were classified in clusters that were irrelevant to the core of the molecules and similar in the R-groups (their composition, geometry, and charge) facing the C-helix and P-loop. The influence of mutations and similarities in pharmacological properties was also observed in different families (Figs. 1b and 2). X-ray crystallography showed that the ATP-binding cleft between the N-lobe and C-lobe can be wider or narrower as a result of different crystal packing [26], which we assume can influence the flexibility of the glycine-rich P-loop. Our

comparative analysis shows highly distinct conformational landscapes for each of the mutant combinations. The most evident cluster of mutants was the triple T790M/L858R/V948R mutant in Family D2 in the N-lobe (Fig. 2) and at average weighted distance of 20.6099 in the whole kinase (Fig. S2 in the supplemental information online).

Most ligands in each family were characterized by a secondary amide (scaffold–NH–R^1) group projecting from the scaffold and bent at the second bond at a torsional angle of 45–75° toward the base of the C-helix. This conforms –R^1 group in a new plane where its aromatic parts (if any) are sandwiched between Leu788 and Thr790 (or Met790 in T790M mutant). In families C, D, and E, the –R^1 group is further extended to make more contacts with Leu777 and Leu788 (Fig. 3c–f), thus shifting the whole ligand molecule backwards (more precisely outwards) with the bent plane sandwiched between Thr790 and Lys745 (instead of Leu788). The narrow nature of that back-pocket has been a major challenge for R-group replacement in the past. Xu et al. [27] described a method where they first screened the –R^1 group for potent ligands targeting the narrow pocket at the base of C-helix, and then screened for ligands selective for EGFR against ERBB2 by modifying the ligand from the opposite side (facing solvent). Several compounds with high potency to EGFR displayed low receptor selectivity. The shift toward mutation selectivity has been a game changer over the past decade. Structural selectivity of gefitinib and erlotinib for the L858R mutant is attributed to their specific recognition of the active kinase state and to weaker ATP binding by L858R EGFR (higher ATP Km value of the mutant EGFR relative to the wild-type). Therefore, the emergence of drug resistance because of the secondary mutation T790M is the result of the restoration of the ATP Km value of the double-mutant enzyme to wild-type levels [28]. An alternative strategy that targets the kinase entrance far from T789M focused on the hydrophobic clamp formed between N-lobe (L718 and V726) and C-lobe (L844). The insight behind this approach was from several EGFR-ligand co-crystals such as 5XDK, 3IKA, and 5X2C [29]. The latest generation of TKIs has no contact or effect on the C-helix.

EGFR is activated through the formation of asymmetric dimer, where the C-lobe of an activator interacts with the N-lobe of a receiver, causing its C-helix to fold inwards (i.e., C-helix-in) [30]. EGFR inhibition via a phenomenon called 'DFG-in-C-helix-out' was observed in the families C2, D1, D2, and E1 and has been previously described particularly for irreversible TKIs [31,32]. The new conformation of a stable outward C-helix results in a larger and more hydrophobic pocket to accommodate an aromatic moiety from the TKI. In the crystal structures 3IKA and 5GNK, the C-helix is pushed farther from the pocket in the latter structure, but forms hydrogen bonds with Asp855, whereas the acrylamide group in the ligands forms a covalent bond with Cys797 [31]. Although the DFG-in-C-helix-out conformation can provide a selective advantage for EGFR kinases, mechanisms that stabilize the C-helix-in conformation would lead to drug resistance [30]. Indeed, oncogenic mutations causing resistant C-helix-in have been reported in HER2, BRAF, and EGFR exon 19 deletions, targeting particularly inhibitors of the C-helix-out families [33]. The role of C-helix rotation in forming inactive conformation has been well studied. The process is mediated by Leu858 (Fig. 1a), which forms a helical turn and hydrophobic interactions with other residues in the N-lobe, thus displacing the C-helix from the active site and rotating it outwards into an inactive conformation [34].

The DFG motif (Asp855–Gly857) in the C-lobe has been targeted by alternative TKI in two different receptors. Covalent ligands were reported to target Cys797 of EGFR L858R in the DFG-in conformation and Cys477 of FGFR4 V550L in the DFG-out conformation [35]. Examples of DFG motif targeting include the crystal structures 4JQ7, 4JQ8, 4JR3, and 4JRV [36].

Concluding remarks and recommendations

In their review, Lionta et al. [5] proposed a protocol for drug design in cases where the target structure exhibits many conformations to: (i) perform RMSD pairwise fitting comparison of the receptor conformations; (ii) perform clustering analysis of fitted models; and (iii) study the changes in the binding site regarding shape

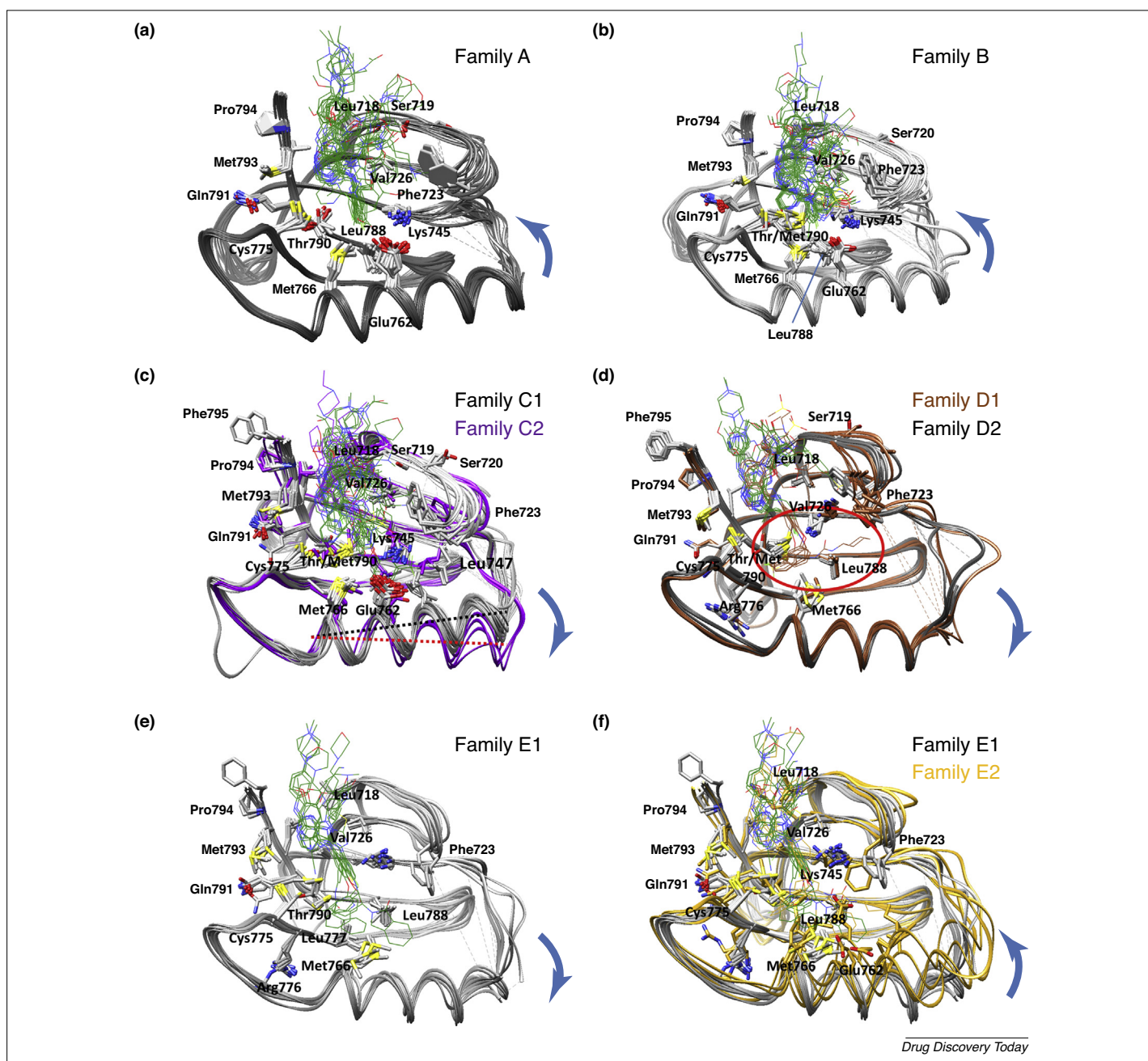Please cite this article in press as:

Features • PERSPECTIVE

FIGURE 3

3D superposition of N-lobe of epidermal growth factor receptor (EGFR) kinase domain (gray ribbons) and complexed ligands (green wire with HETATOM coloring). Residues within a 3-Å distance are shown as gray sticks with HETATOM coloring. Subfamilies are colored accordingly. **(a)** Family A (22 structures with one or no mutations in EGFR) showing the C-helix in proximity to ligands (arrow) and contacts with Glu762. **(b)** Family B (31 structures with mostly double EGFR mutations) showing T790M mutants and the C-helix in proximity to ligands (arrow) by contacts with Glu762. **(c)** Family C1 in gray (21 structures with one or two EGFR mutations) showing T790M mutants and the C-helix in proximity to ligands by contacts with Glu762. Family C2 in purple (four structures) showing shifting of the C-helix away from ligands because of a steric effect. **(d)** Family D1 in brown (seven wild-type EGFR structures) showing steric effect of ligands (brown wire inside the red circle) and the C-helix is shifting away with no contacts against Glu762. Family D2 in gray (ten triple-mutated EGFR structures) showing shifting of the C-helix away from ligands possibly because of effects of mutations. **(e)** Family E1 in gray (ten structures) showing the steric effect of ligands and the C-helix shifting away with no contacts against Glu762. (F) Family E1 and E2 superposed (in gray and gold, respectively). Family E2 (five highly divergent structures) showing a back-shift in the C-helix with ligand contacts against Glu762 and Ile759. Visualization of protein and ligand 3D Structures was performed using UCSF Chimera (version 1.10.2). The matchmaker plugin was used for superposition of all heavy atoms via the BLOSUM-62 scoring matrix and Needleman-Wunsch alignment algorithm.

and volume and use it to filter out unsuitable ligands from the library. We suggest that it is important to clarify some peculiarities in using RMSD in comparative modeling. RMSD, which is commonly used among modelers, can provide a biased estimation of model similarity particularly when an entire subdomain in the structure (such as the case of C-helix in EGFR kinase) is shifted without complete changes inside that subdomain. To solve this bias, RMSD will have to be limited to isolated fragments (of local regions) of the structure [37], or instead, a score

Please cite this article in press as:

that considers local regions and includes local/global fitting can be used, such as the global distance test (GDT) [38]. Another widely accepted alternative is the template modeling scores (TM-score and TM-align), which, in additional to local/global fitting, take the length of polypeptide into consideration [39]. We provide the values for the TM-score and GDT scores of the N-lobe fittings of EGFR structures in Table S4 in the supplemental information online.

Here, we have clarified some of the confusion regarding EGFR structure-based design and have provided a rationale for the applications for which these structure families can be used. For example, 3D structures of the A, B, and C1 families are C-helix-in with two mutations or less. They are suitable for all three generations of TKIs with minimum interactions with the C-helix. The remaining families are C-helix-out conformations with three or less mutations per structure, which are most suitable for targeting of the hydrophobic pocket formed by the C-helix-out conformation.

Declaration of Competing Interest

The authors report no declarations of interest.

Acknowledgments

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at https://doi.org/10.1016/j.drudis.2020.10.007.

References

1 Verlinde, C.L.M.J. and Hol, W.G.J. (1994) Structure-based drug design: progress, results and challenges. *Structure* 2, 577–587
2 Kanakaveti, V. *et al.* (2020) Quantitative structure-activity relationship in ligand based drug design: concepts and applications. In *Protein interactions: computational methods, analysis and applications* (Gromiha, M.M., ed.), pp. 333–349, World Scientific
3 Sledz, P. and Caflisch, A. (2018) Protein structure-based drug design: from docking to molecular dynamics. *Curr Opin Struct Biol* 48, 93–102
4 Engel, J. *et al.* (2016) Hope and disappointment: covalent inhibitors to overcome drug resistance in non-small cell lung cancer. *ACS Med Chem Lett* 7, 2–5
5 Lionta, E. *et al.* (2014) Structure-based virtual screening for drug discovery: principles, applications and recent advances. *Curr Top Med Chem* 14, 1923–1938
6 Cheng, T. *et al.* (2012) Structure-based virtual screening for drug discovery: a problem-centric review. *AAPS J* 14, 133–141
7 Langdon, S.R. *et al.* (2010) Bioisosteric replacement and scaffold hopping in lead generation and optimization. *Mol Inform* 29, 366–385
8 Lo, Y.-C. *et al.* (2018) Machine learning in chemoinformatics and drug discovery. *Drug Discov Today* 23, 1538–1546
9 Haddad, Y. *et al.* (2019) Rotamer dynamics: analysis of rotamers in molecular dynamics simulations of proteins. *Biophys J* 116, 2062–2072
10 Smith, M. and Smith, J.C. (2020) Repurposing therapeutics for COVID-19: supercomputer-based docking to the SARS-CoV-2 viral spike protein and viral spike protein–human ACE2 interface. *ChemRxiv* 11, 2020. http://dx.doi.org/10.26434/chemrxiv.11871402.v4 [published online March]
11 Buonerba, C. *et al.* (2019) Predictors of outcomes in patients with EGFR-mutated non-small cell lung cancer receiving EGFR tyrosine kinase inhibitors: a systematic review and meta-analysis. *Cancers* 11, 1–18
12 Leonetti, A. *et al.* (2019) Resistance mechanisms to osimertinib in EGFR-mutated non-small cell lung cancer. *Br J Cancer* 121, 725–737
13 Tan, C.-S. *et al.* (2018) Third generation EGFR TKIs: current data and future directions. *Mol Cancer* 17, 1–14
14 Lategahn, J. *et al.* (2019) Inhibition of osimertinib-resistant epidermal growth factor receptor EGFR-T790M/C797S. *Chem Sci* 10 (46), 10789–10801
15 Jia, Y. *et al.* (2016) Overcoming EGFR(T790M) and EGFR (C797S) resistance with mutant-selective allosteric inhibitors. *Nature* 534 (7605), 129–132
16 Wang, S. *et al.* (2017) EAI045: The fourth-generation EGFR inhibitor overcoming T790M and C797S resistance. *Cancer Lett* 385, 51–54
17 To, C. *et al.* (2019) Single and dual targeting of mutant EGFR with an allosteric inhibitor. *Cancer Discov* 9, 926–943
18 Oronsky, B. *et al.* (2018) Navigating the 'No Man's Land' of TKI-failed EGFR-mutated non-small cell lung cancer (NSCLC): a review. *Neoplasia* 20, 92–98
19 Rebuzzi, S.E. *et al.* (2020) Combination of EGFR-TKIs and chemotherapy in advanced EGFR mutated NSCLC: Review of the literature and future perspectives. *Crit Rev Oncol Hematol* 146, 1–10
20 Le Tourneau, C. *et al.* (2019) Molecular profiling in precision medicine oncology. *Nat Med* 25, 711–712
21 Guardiola, S. *et al.* (2019) A third shot at EGFR: new opportunities in cancer therapy. *Trends Pharmacol Sci* 40, 941–955
22 Chen, Y. *et al.* (2020) Discovery of new thieno 3,2-d pyrimidine derivatives targeting EGFR(L858R/T790M) NSCLCs by the conformation constrained strategy. *Eur J Med Chem* 199, 1–13
23 Karnik, K.S. *et al.* (2020) Development of triple mutant T790M/C797S allosteric EGFR inhibitors: a computational approach. *J Biomol Struct Dyn* . http://dx.doi.org/10.1080/07391102.2020.1786460 [published online July 1, 2020]
24 Sun, X.Q. *et al.* (2014) Structure-based ensemble-QSAR model: a novel approach to the study of the EGFR tyrosine kinase and its inhibitors. *Acta Pharmacol Sin* 35, 301–310
25 Zhou, W. *et al.* (2009) Novel mutant-selective EGFR kinase inhibitors against EGFR T790M. *Nature* 462 (7276), 1070–1074
26 Kong, L.-L. *et al.* (2017) Structural pharmacological studies on EGFR T790M/C797S. *Biochem Biophys Res Commun* 488, 266–272
27 Xu, G. *et al.* (2008) Discovery of novel 4-amino-6-arylaminopyrimidine-5-carbaldehyde oximes as dual inhibitors of EGFR and ErbB-2 protein tyrosine kinases. *Bioorg Med Chem Lett* 18, 3495–34999
28 Gajiwala, K.S. *et al.* (2013) Insights into the aberrant activity of mutant EGFR kinase domain and drug recognition. *Structure* 21, 209–219
29 Zhu, S.-J. *et al.* (2018) Structural insights into drug development strategy targeting EGFR T790M/C797S. *Oncotarget* 9 (17), 13652–13665
30 Novotny, C.J. *et al.* (2016) Overcoming resistance to HER2 inhibitors through state-specific kinase binding. *Nat Chem Biol* 12, 923–930
31 Wang, A. *et al.* (2017) Discovery of (R)-1-(3-(4-amino-3-(3-chloro-4-(pyridin-2-ylmethoxy) phenyl)-1 H-pyrazolo [3,4-d] pyrimidin-1-yl) piperidin-1-yl) prop-2-en-1-one (CHMFL-EGFR-202) as a novel irreversible EGFR mutant kinase inhibitor with a distinct binding mode. *J Med Chem* 60, 2944–2962
32 Hu, C. *et al.* (2017) Discovery and characterization of a novel irreversible EGFR mutants selective and potent kinase inhibitor CHMFL-EGFR-26 with a distinct binding mode. *Oncotarget* 8, 18359–18372
33 Foster, S.A. *et al.* (2016) Activation mechanism of oncogenic deletion mutations in BRAF, EGFR, and HER2. *Cancer Cell* 29, 477–493
34 Yun, C.-H. *et al.* (2007) Structures of lung cancer-derived EGFR mutants and inhibitor complexes: mechanism of activation and insights into differential inhibitor sensitivity. *Cancer Cell* 11, 217–227
35 Tan, L. *et al.* (2014) Development of covalent inhibitors that can overcome resistance to first-generation FGFR kinase inhibitors. *Proc Natl Acad Sci U S A* 111 (45), 4869–4877
36 Peng, Y.-H. *et al.* (2013) Protein kinase inhibitor design by targeting the Asp-Phe-Gly (DFG) motif: the role of the DFG motif in the design of epidermal growth factor receptor inhibitors. *J Med Chem* 56, 3889–3903
37 Haddad, Y. *et al.* (2020) Ten quick tips for homology modeling of high-resolution protein 3D structures. *PLoS Comp Biol* 16, 1–19
38 Zemla, A. (2003) LGA: a method for finding 3D similarities in protein structures. *Nucleic Acids Res* 31 (13), 3370–3374
39 Zhang, Y. and Skolnick, J. (2005) TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res* 33, 2302–2309

Yazan Haddad[1,2]
Marek Remes[1,2]
Vojtech Adam[1,2]
Zbynek Heger[1,2,*]

[1]*Department of Chemistry and Biochemistry, Mendel University in Brno, Zemedelska 1, CZ-613 00 Brno, Czech Republic1 Department of Chemistry and Biochemistry, Mendel University in Brno, Zemedelska 1, CZ-613 00 Brno, Czech Republic* [2]*Central European Institute of Technology, Brno University of Technology, Purkynova 656/123, 612 00 Brno, Czech Republic2 Central European Institute of Technology, Brno University of Technology, Purkynova 656/123, 612 00 Brno, Czech Republic*

*Corresponding author:.

Please cite this article in press as: