MedChemComm

Accepted Manuscript

ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/medchemcomm

Concise Article

Target-based Analysis of Ionization States of Bioactive Compounds

Shilva Kayastha[#], Antonio de la Vega de León[#], Dilyana Dimova, and Jürgen Bajorath[*]

Department of Life Science Informatics, Bonn-Aachen International Center for Information Technology, Rheinische Friedrich-Wilhelms-Universität Bonn, Dahlmannstr. 2, D-53113 Bonn, Germany.

[#]The contributions of these authors should be considered equal.

*To whom correspondence should be addressed:

Tel: +49-228-2699-306, Fax: +49-228-2699-341, E-mail: bajorath@bit.uni-bonn.de

Summary

A systematic analysis of ionization states of current bioactive compounds is presented. Ionization states were related to biological activities on the basis of high-confidence activity data. The majority of bioactive compounds were found to be basic or neutral under physiological conditions. In addition, chemical neighborhoods of active compounds frequently contained analogs with different ionization states that were activity-conservative. However, a variety of targets were identified that displayed clear preferences for specific ionization states in compounds active against them. In this context, notable differences in the distribution of ionization states were detected for compounds active against different target superfamilies. Furthermore, under physiological pH, differences in ionization states of active compounds were tolerated by many targets. However, in a number of instances, ionization states of highly and weakly potent compounds active against the same target were found to be distinct, providing guidelines for compound design and optimization.

Introduction

The charge state of small molecules is a major determinant of biological activity and drug action.[1-5] It has been estimated that the majority of drugs are partly ionized under physiological conditions.[1] A convincing perspective has also been provided on the critical role compound ionization states play at different stages of pharmaceutical development.[3] Importantly, differences in the pH in various cellular compartments or extracellular environments can modulate ionization states of active compounds, alter their properties *in vivo*, and affect pharmacological profiles. A refined charge state profile of oral drugs indicated that nearly 80% of them contained ionizable groups, while only ~12% were neutral.[3] In addition, acid/base profiles of drugs directed against major target classes (including proteases, kinases, G protein coupled receptors, and various ion channels) were studied and notable differences between these profiles were identified as well as differences between individual target families comprising a given class.[4] Furthermore, the ionization states of drugs and screening compounds were compared. It was found that drugs contained a much higher proportion of both carboxylic acid groups and aliphatic amines than compounds from various sources available for screening, indicating that many compounds in screening collections might lack relevance for drug discovery, given the prevalent charge states of drugs.[4]

In a recent extensive analysis of publicly available compound data,[5] acidic and basic bioactive compounds and drugs were compared and the influence of ionization states on a variety of calculated or observed physico-chemical and pharmacological properties was studied. For this purpose, acids and bases were classified as compounds that were proton donors and acceptors, respectively, and at least 50% ionized under physiological pH of 7.4 (calculated using the

Henderson-Hasselbalch equation[6]). Major conclusions from this work included that weak bases containing N-heterocycles are frequent among drugs and that their physico-chemical and pharmacological properties are by and large tolerable, that strongly basic compounds should best be avoided due to unfavourable properties, and that acids are under-represented in drugs but should merit further consideration.[5] In addition to their thorough analysis of drugs, Charifson and Walters also analyzed the activity distribution of bioactive compounds with different ionization states across cell-based assays and of compounds tested in at least 20 assays. It was found that acidic compounds were generally less active in cellular assays than compounds with other charge states and that basic compounds were overall less selective than acidic or neutral ones.[5]

Herein we also report a large-scale analysis of bioactive compounds with respect to ionization states, albeit with different focal points. Our analysis primarily focuses on relationships between compound ionization states, structural similarity, and potency and exclusively uses high-confidence activity data.

Methods

Small molecules can generally be classified according to ionization states as bases, acids, neutral compounds, or zwitterionic molecules.[1,5] The dissociation constant (K_a) is an equilibrium constant determining ionization states. Commonly used is the logarithmic form of the dissociation constant (pK_a), defined as the negative decadic logarithm of K_a ($-\log_{10} K_a$). To account for acidic and basic properties of small molecules, two different pK_a types are considered including the acidic pK_a (A_pK_a) and the basic pK_a (B_pK_a). Following this distinction, A_pK_a is defined as the pK_a for the most acidic group in a given molecule whereas B_pK_a is defined as the pK_a for the most basic group. For all compounds analyzed herein, calculated values of A_pK_a and B_pK_a were extracted from the ChEMBL database[7] (version 19).

Compounds were assigned to four ionization state classes (IS-classes) including basic, acidic, neutral, and zwitterionic compounds on the basis of A_pK_a and B_pK_a values relative to the physiological pH of 7.4, following the approach of Charifson and Walters.[5] Accordingly, compounds with an acidic or basic group were classified as acids or bases, respectively, if they were more than 50% ionized at pH 7.4. In addition, compounds containing acidic and basic groups were classified as acids if the acidic group was more than 50% ionized and the basic group less than 50%, as bases if the basic group was more than 50% ionized and the acidic group less than 50%, and as zwitterionic compounds if both groups were ionized more than 50%. Furthermore, compounds were classified as neutral if acidic and/or basic groups were both

ionized less than 50% under physiological pH. If A_Pk_a and B_pK_a values were not available for a compound, it was not assigned (NA).

From ChEMBL (version 19), compounds active against human targets at the highest confidence level (confidence score 9) were extracted for which assay-independent equilibrium constants (K_i values) were available as potency measurements. Compounds with multiple measurements for the same target were only considered if all values fell within the same order of magnitude. Then the geometric mean of these was calculated as the final potency annotation. If multiple stereoisomers of a compound with potency within one order of magnitude were available, the compound was retained. All qualifying compounds were organized in individual activity classes (target sets). A total of 719 K_i-based target sets were obtained comprising 80,776 compounds.

To assess structural relationships between active compounds, matched molecular pairs (MMPs)[8] were calculated. MMPs consist of pairs of compounds that are only distinguished by a structural change at a single site (chemical transformation).[8,9] Size restrictions were introduced to limit transformations to small structural modifications.[10] Accordingly, the size (number of heavy atoms) of the shared MMP core had to be at least twice the size of each of the exchanged substructures. In addition, the size of each transformation fragment was limited to a maximum of 13 heavy atoms and the difference between the exchanged fragments to eight heavy atoms.[10] For each target set, transformation size-restricted MMPs were systematically calculated using an in-house implementation of the algorithm by Hussain and Rea[8] utilizing the *OEChem* toolkit.[11] MMPs involving NA compounds were omitted from further analysis. Furthermore, target sets yielding fewer than 50 MMPs were excluded. A total of 338,419 MMPs were obtained that exclusively involved a total of 66,871 IS-class compounds from 290 different target sets.

For each classified compound, its chemical neighborhood was determined by combining all of its MMP partners (structural analogs) within a target set and three neighborhood categories (CATs) were defined as follows: (I) all neighbors, (II) only a subset of neighbors, or (III) none of the neighbors belonged to the same IS-class as the reference compound.

Results and discussion

Compound ionization state class distribution

Figure 1 reports the IS-class assignment for all qualifying bioactive compounds with available high-confidence activity data (only 8.4% of all compounds could not be assigned to one of the four IS-classes, due to missing pK_a values). Consistent with previous findings that many drugs are weak bases under physiological conditions, we also determined that bases were prevalent among bioactive compounds (39.2%). Interestingly, however, a comparable proportion of bioactive compounds was neutral (38.6%), regardless of their activity. Considering the entire potency range, only 3.5% and 10.3% of active compounds were zwitterionic and acidic, respectively. The observed global distribution over IS-classes was essentially mirrored by a subset set of 39,783 compounds with a potency of at least 100 nM (with relative class deviations < 2%).

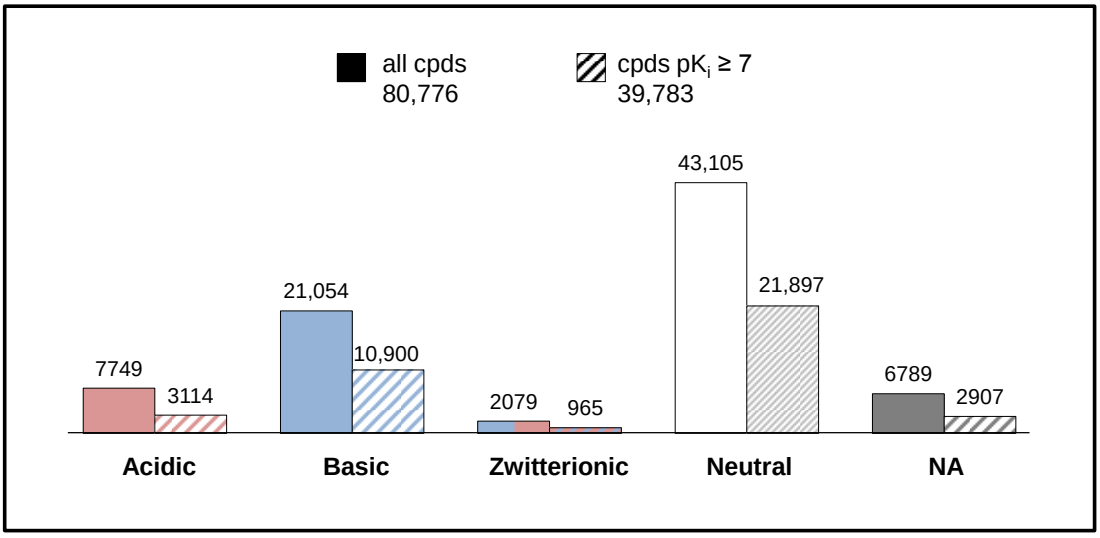

Figure 1. Ionization state class distribution. Reported is the class distribution for all 80,776 qualifying compounds (solid bars) and a subset of 39,783 compounds with a potency of at least 100 nM or higher (striped bars).

Chemical neighborhood analysis

We then systematically explored the chemical neighborhoods of compounds in all IS-classes through MMP calculations. The majority of MMPs (86.4%) were formed between compounds belonging to the same IS-class. Hence, most structural analogs of classified compounds had conserved ionization sates. However, many compounds had at least one or a few structural analogs belonging to a different IS-class. We found that 68.8% of all neighborhoods consisted of compounds with conserved ionization states, while 28.7% of the neighborhoods contained one or more compounds belonging to a different IS-class than the reference molecule. In addition, in 2.5% of the neighborhoods, all compounds belonged to IS-classes different from the reference molecule. Hence, about one third of all neighborhoods were heterogeneous in their IS-class composition. However, these frequently occurring differences in ionization states were activity-conservative. **Figure 2** shows exemplary compound neighborhoods of different composition.

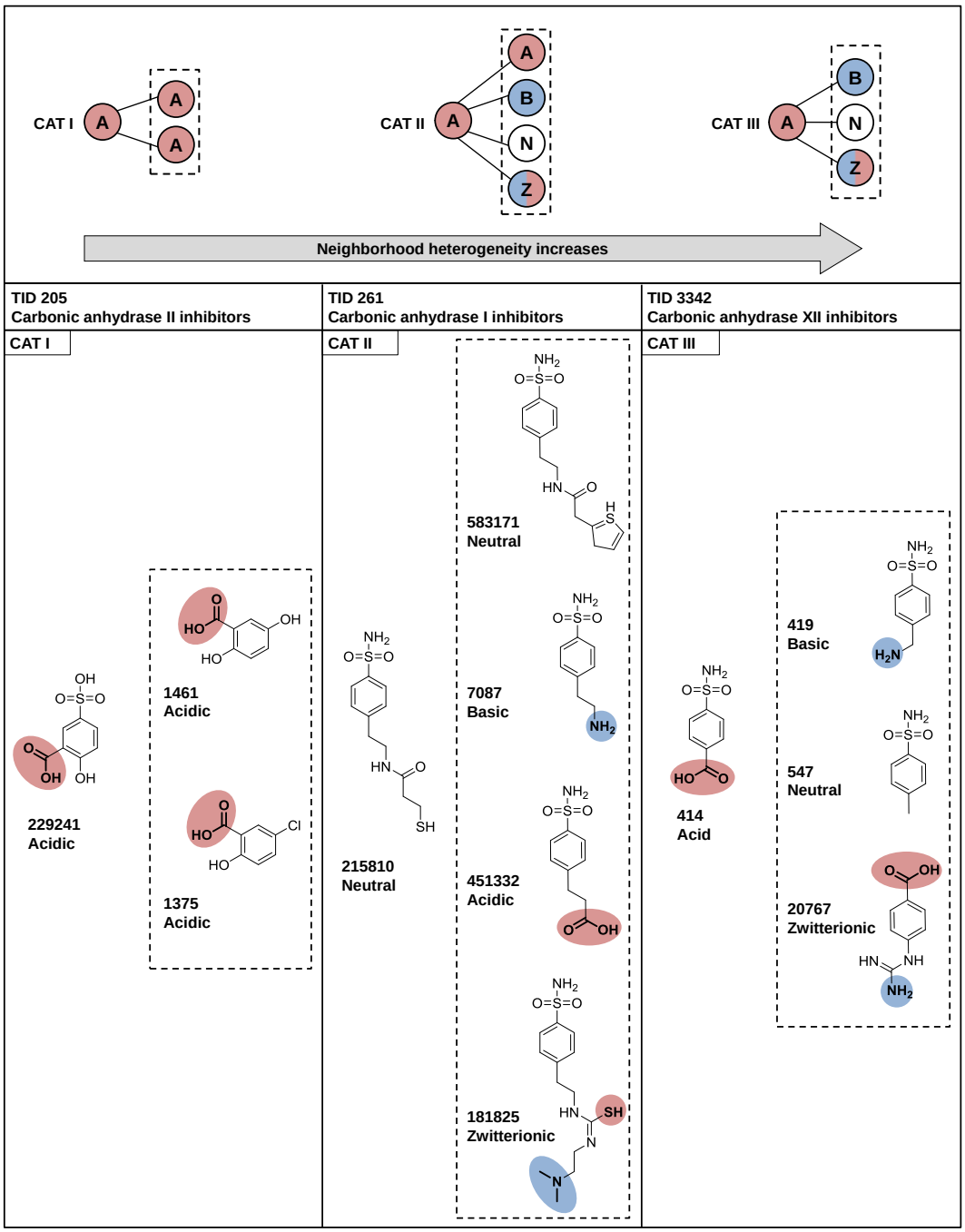

Figure 2. Chemical neighborhoods. The figure shows the IS-class composition (red, acidic; blue, basic; white, neutral; dual colored, zwitterionic) of exemplary chemical neighborhoods of category (CAT) I-III with inhibitors of different carbonic anhydrase isoforms (TIDs report

ChEMBL target set IDs). Functional groups (acidic, red; basic, blue) ionized at physiological pH are depicted in bold and highlighted. ChEMBL compound ID and IS-classes are given.

Ionization state class distribution over activity classes and target superfamilies

We next determined the distribution of IS-classes over target sets. Target sets with fewer than 10 compounds or more than 20% unclassified (NA) compounds were excluded from this analysis. In nearly 90% of 351 qualifying target sets, more than half of the compounds belonged to the same IS-class and in 40%, more than 80% belonged to the same class. Hence, although compound neighborhoods were frequently found to be heterogeneous in their ionic state composition, as discussed above, many target sets displayed a strong ionization state preference. In most cases, basic or neutral compounds dominated. This can also be seen in **Table 1** that reports the top 20 target sets (comprising at least 200 compounds) having the highest percentage of compounds belonging to the same IS-class. The ranking contains many different G protein coupled receptors (GPCRs), but also transporters and proteases. In addition to basic compounds, strong preferences for neutral (e.g., vanilloid receptor ligands) and acidic compounds (e.g., prostaglandin D2 receptor 2 ligands) were also observed.

Table 1. Target sets with ionization state class dominance.

Target ID	Target name	# cpds	Dominant IS-class
5071	Prostaglandin D2 receptor 2	468	99% acidic
4794	Vanilloid receptor	253	97% neutral
259	Melanocortin receptor 4	1217	92% basic
264	Histamine H3 receptor	2023	92% basic
1898	Serotonin 1b (5-HT1b) receptor	364	92% basic
335	Protein-tyrosine phosphatase 1B	243	91% acid
344	Melanin-concentrating hormone receptor 1	846	90% basic
4644	Melanocortin receptor 3	350	90% basic
4608	Melanocortin receptor 5	268	88% basic
1983	Serotonin 1d (5-HT1d) receptor	359	87% basic
1800	Corticotropin releasing factor receptor 1	473	84% neutral
222	Norepinephrine transporter	1010	84% basic
232	Alpha-1b adrenergic receptor	290	84% basic
228	Serotonin transporter	1337	83% basic
2492	Neuronal acetylcholine receptor protein alpha-7 subunit	253	83% basic
238	Dopamine transporter	867	81% basic
3798	Calcitonin gene-related peptide type 1 receptor	349	81% neutral
1916	Alpha-2c adrenergic receptor	295	80% basic
2954	Cathepsin S	375	80% neutral
210	Beta-2 adrenergic receptor	241	80% basic

The top 20 target sets (with ChEMBL IDs) with highest percentages of compounds belonging to the same IS-class are reported.

Target sets were also organized into superfamilies and the IS-class distribution of their ligands was determined, as reported in **Figure 3**. Clear trends were observed. For example, 50% of available enzyme inhibitors and 63% of membrane receptor ligands were neutral. While enzyme inhibitors displayed a balanced distribution of acidic (15%) and basic (20%) compounds, membrane receptor ligands showed a notable preference for basic (46%) over acidic compounds (8%). Furthermore, 67% of all compounds active against transporters were bases. Moreover, 50% and 26% of ion channel ligands were basic and neutral compounds, respectively, whereas only 3% of them were acids. Thus, there were marked differences in ionization state preferences for compounds active against different target superfamilies.

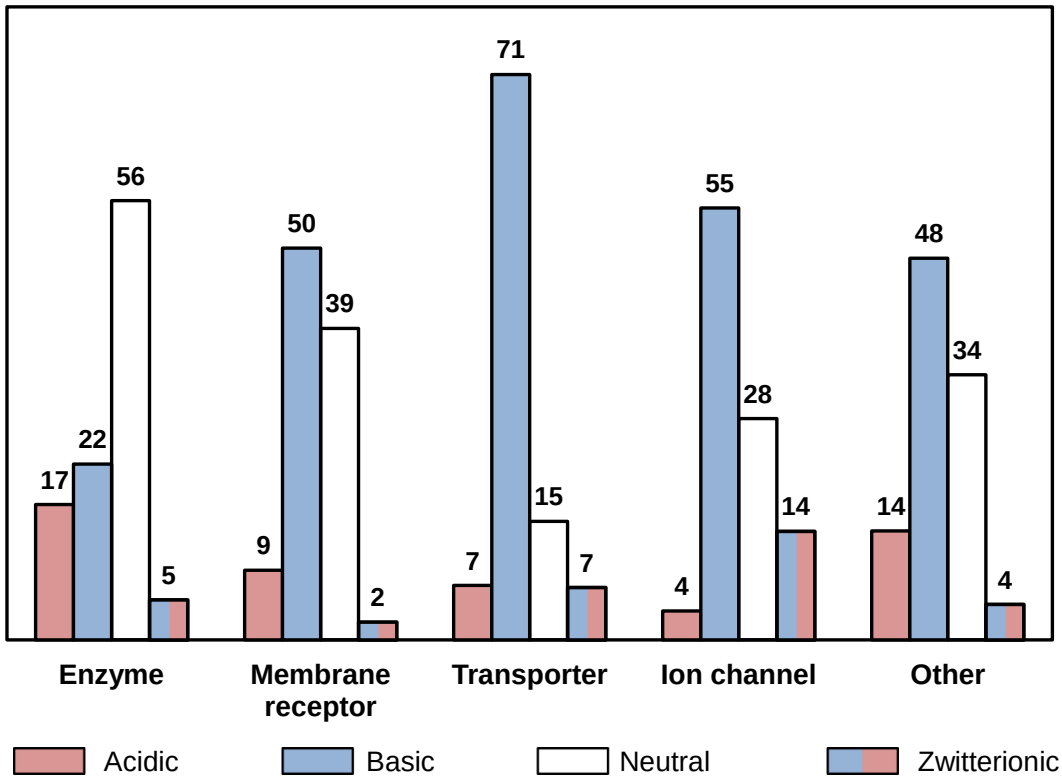

Figure 3. Compound IS-class distribution over target superfamilies. Reported is the IS-class distribution (red, acidic; blue, basic; white, neutral; dual colored, zwitterionic) of ligands of four target superfamilies (plus "Other"). For each superfamily, the percentage of active compounds belonging to each category is given.

Potency range distribution of ionization state classes

Finally, the potency range distribution of IS-classes was studied in detail for all target sets. Although there were no significant differences between the global IS-class distributions of all bioactive compounds and a subset of highly potent compounds, as reported above (and shown in **Figure 1**), we detected 57 target sets with notable differences in IS-class distributions between weakly (pKi $\leq$ 6; WP) and highly potent (pKi $\geq$ 7; HP) compounds. These 57 target sets belonged to three superfamilies (enzymes, membrane receptors and transporters). The majority of the targets belonged to enzyme inhibitors (28) followed by membrane receptors (19). Only two targets were transporters. The IS class distribution of HP and WP compounds in these target sets displayed significant differences. In 27 enzyme inhibitor sets, more than 80% of the HP compounds were zwitterionic (and less than 20% of WP compounds were zwitterionic), whereas for the majority of membrane receptors, the percentage of HP zwitterionic compounds was less than 20%. The structures of the HP and WP compound sets for specific IS-classes and superfamilies were mostly distinct, because the overlap in scaffolds[12] between these sets of compounds rarely exceeded 5%. In addition, the sets of HP and WP compounds had high intra-set diversity, because each scaffold represented on average only one to two different compounds. In **Figure 4**, representative examples are shown. **Figure 4A** reports the IS-classes of neurokinin 2 receptor antagonists. Among the highly potent ligands, there was a clear preference for basic

over neutral compounds, whereas the trend was reversed for weakly potent compounds where neutral species were found to dominate. In **Figure 4B**, a notable enrichment of basic compounds among weakly potent urokinase-type plasminogen activators is observed. In this case, highly potent compounds had different ionization states (which was rather unusual). Furthermore, **Figure 4C** shows a reversal in the distribution of acidic and neutral compounds among highly and weakly potent inhibitors of inosine-5' monophosphate dehydrogenase 2, corresponding to observations made for basic and neutral compounds in **Figure 4A**. Furthermore, **Figure 4D** shows that basic compounds were frequently observed among weakly potent coagulation factor XI inhibitors, whereas highly potent inhibitors were zwitterionic, without an exception.

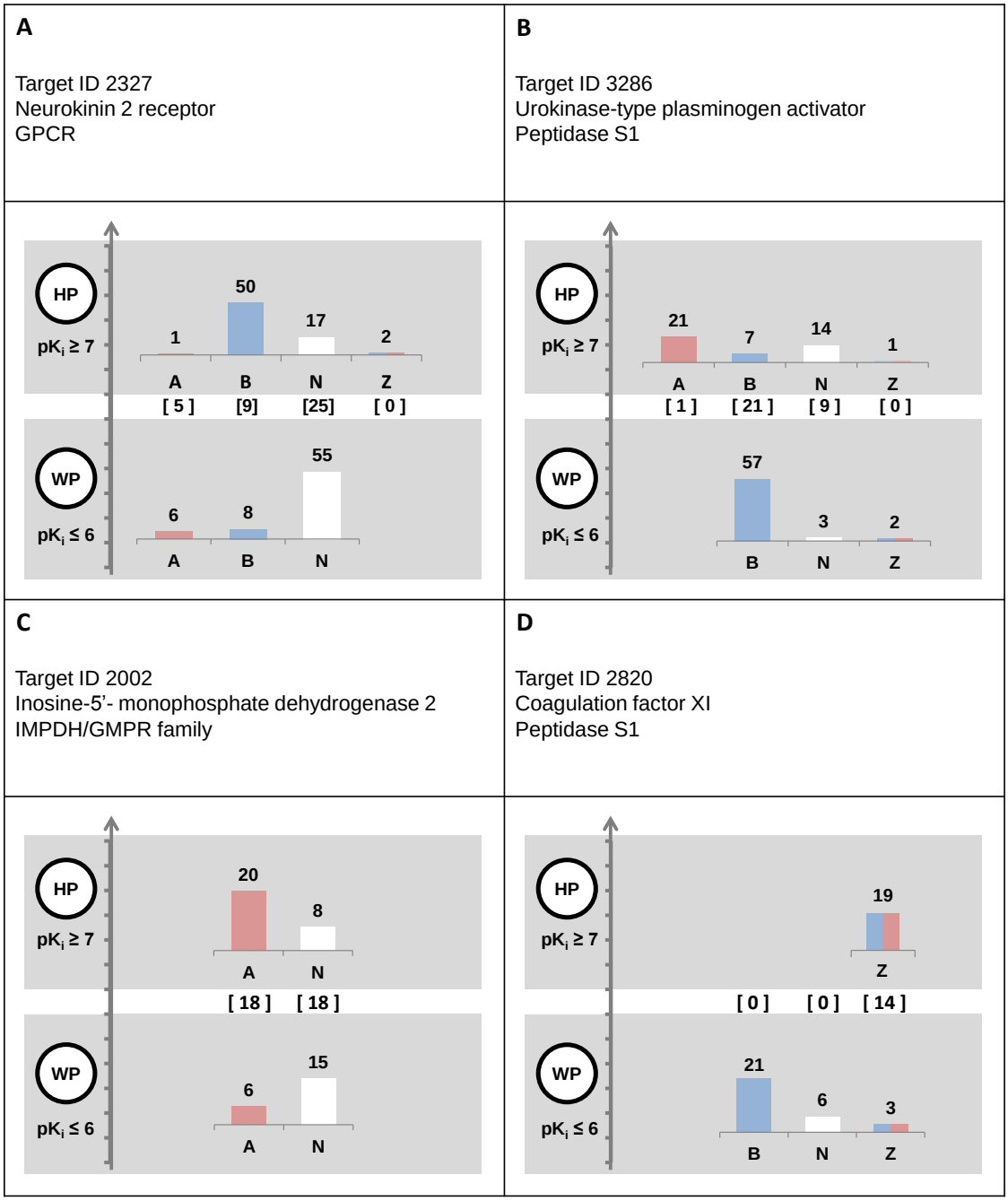

Figure 4. IS-class changes over potency ranges. For four exemplary target sets (with targets belonging to different families), the IS-class distribution (red, acidic; blue, basic; white, neutral; dual colored, zwitterionic) for highly potent (HP, $pK_i \geq 7$) and weakly potent (WP, $pK_i \leq 6$)

compounds is shown. The number of compounds in different IS-classes falling into the intermediate potency interval is also given.

Conclusions

We have carried out a large-scale analysis of calculated ionization states in bioactive compounds and their distribution across different targets and families that complements and further extends previous investigations. Ionization states in chemical neighborhoods of bioactive compounds were determined across different potency ranges, setting our analysis apart from previous studies. Furthermore, different from earlier studies that strongly (but not exclusively) focused on drugs, we comprehensively analyzed currently available spectrum of bioactive compounds and exclusively based our analysis on carefully selected high-confidence activity data. Our results reveal the presence of an uneven global distribution of ionization states across the bioactive compounds, the majority of which were basic or neutral under physiological conditions. Individual target sets were found to display significant differences in preferred ionization states. Similar observations were made for different target superfamilies. Systematic MMP analysis revealed that changes in ionization states frequently occurred among structural analogs. Moreover, potency range-dependent differences in the distribution of ionization states were detected in a variety of target sets. We found that ionization states of highly potent compounds were often different from weakly potent ones. In a number of cases, weakly potent compounds were predominantly basic, while different ionization states were observed among highly potent ones. In other instances, weakly potent compounds were mostly neutral, whereas highly potent compounds were charged. Thus, for a variety of targets, preferred ionization states characteristic of highly potent compounds can be identified. The presence of preferred ionization states in highly potent compounds for different targets provides valuable guidelines for compound design and optimization.

Acknowledgement

We are grateful to OpenEye Scientific Software, Inc., for the free academic license of the OpenEye Toolkits. We thank Norbert Furtmann for helpful discussions.

References

1. D.T. Manallack, *Perspect. Medicin. Chem.* 2007, **1**, 25–38.

2. M.P. Gleeson, *J. Med. Chem.* 2008, **51**, 817–834.

3. D.T. Manallack, R.J. Prankerd, E. Yuriev, T.I. Oprea and D.K. Chalmers, *Chem. Soc. Rev.* 2013, **42**, 485–496.

4. D.T. Manallack, R.J. Prankerd, G.C. Nassta, O. Ursu, T.I. Oprea and D.K. Chalmers, *ChemMedChem* 2013, **8**, 242–255.

5. P.S. Charifson and W.P. Walters, *J. Med. Chem.* 2014, **57**, 9701–9717.

6. H.N. Po and N.M. Senozan, *J. Chem. Educ.* 2001, **78**, 1499–1503.

7. A. Gaulton, L.J. Bellis, A.P. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, D. Michalovich, B. Al-Lazikani and J.P. Overington, *Nucleic Acids Res.* 2012, **40**, D1100–1107.

8. E. Griffen, A.G. Leach, G.R. Robb and D.J. Warner, *J. Med. Chem.* 2011, **54**, 7739–7750.

9. J. Hussain and C. Rea, *J. Chem. Inf. Model*. 2010, **50**, 339–348.

10. X. Hu, Y. Hu, M. Vogt, D. Stumpfe and J. Bajorath, *J. Chem. Inf. Model.* 2012, **52**, 1138–1145.

11. OpenEye Scientific Software Inc: Santa Fe, NM.

12. G. W. Bemis and M. A. Murcko. *J. Med. Chem.* 1996, **39**, 2887–2893.

Graphical Contents Entry

Ionization states within a chemical neighborhood. Shown are an acidic and a basic analog of a neutral compound.