Correction to Ligands and Receptors with Broad Binding Capabilities Have Common Structural Characteristics: An Antibiotic Design Perspective

György Abrusán* and Joseph A. Marsh

J. Med. Chem. 2019, 62 (21), 9357–9374. DOI: 10.1021/acs.jmedchem.9b00220

Pages 9357 (Abstract), 9363, and 9370. We erroneously conclude that halogen-containing fragments are significantly enriched in the compounds that can bind several receptors.

Page 9364 and Supporting Information. Figure 6, panels B–D, and Supporting Information Figure S5, panels B–D, and the corresponding figure legends indicate a significant enrichment of halogen containing fragments that disappears when the number of heavy atoms is used as the factor controlling for size. See Figure 3 below for the same panels calculated using the number of heavy atoms.

The erroneous conclusion of our analysis was caused by a feature of the latent space of the variational autoencoder (VAE) that we overlooked. The latent space of the VAE we used (and VAEs in general) is highly structured, where similar chemical compounds are generally located close to each other. However, their location is also strongly influenced by the length of their SMILES string. Figure 1A in this Addition/Correction shows the distance from origin (norm) of the 26,000 de novo designed ligands from our study. Compounds with shorter SMILES are located closer to the origin and have a smaller norm (Figure 1B and Figure 1C; lines represent loess regressions for compounds with and without heavy halogens). Since in our article we used the distance in the latent space as a measure of chemical similarity, when one controls for compound size with molecular weight, this results in an overrepresentation of compounds with heavy elements, as, due to their shorter SMILES, they are located somewhat closer to each other in the latent space. This is not the case when the number of heavy atoms is used, at least when the smallest compounds are excluded. (Note that this effect is generally not limited to halogens.) Improving the latent space of autoencoders is an

Figure 1. Correlations between compounds size and the norm (distance from the origin) of all 26,000 de novo designed ligands in the VAE latent space. The strongest relationship is with SMILES length. Lines represent loess regressions for compounds with and without heavy halogens.

Figure 2. Structural similarity between selected compounds based on number of heavy atoms and molecular weight. Tanimoto similarity was calculated with Open Babel using the FP4 fragment set.

Published: December 13, 2019
area of active research, and since the publication of the VAE\textsuperscript{1} by the Aspuru-Guzik lab that we used, VAEs with different architectures have been developed (see ref 2 for an overview). As most of the newer VAE architectures also rely on SMILES as input, these characteristics of the VAE suggest that unless molecular weight is a feature included in the training, in the case of SMILES based VAEs one should use the number of heavy atoms, rather than molecular weight, when navigating their latent space.

To test for the consequences of this effect, we repeated the fragment enrichment analysis (Figures 6 and S5 of the article), using the number of heavy atoms as the controlling factor for size. With a similar distance cutoff of 13, compounds with 21–25 heavy atoms resulted in approximately similar numbers of chemical compounds with neighbors in at least two additional species: 66 in de novo set 1 and 69 in set 2. (Above 25 heavy atoms there are hardly any compounds with distance smaller than 13.) The structures obtained when using heavy atoms as

\textbf{Figure 3.} Fragment enrichment of the ligands sets using the number of heavy atoms for controlling ligand size. Panels A, B, and C correspond to panels B, C, and D of Figure 6 of the article. Panels D, E, and F correspond to panels B, C, and D of Supporting Information Figure S5 of the article. (A–C) Abundance, enrichment and 2D structures of the significantly enriched fragments in set 1. (D–F) The same for set 2.
the controlling parameter are highly similar to the ones reported in the article, where mol. weight was used as the factor controlling for size (∗≥300 Da, which in practice resulted in mol. weight ranges of 300−440 Da and 300−370 Da). The majority of them have Tanimoto score 1 with its best matching structure (Figure 2), and most compounds that are not identical have scores above 0.7.

The fragment enrichment analysis of the two new ligand sets (Figure 3 in this Addition/Correction) indicates that the enriched fragments obtained by controlling size with heavy atoms are qualitatively similar to the ones obtained by controlling with mol. weight: carboxyl (sid.23), morpholine (sid.68), quinazolinone (lnk.308, lnk.333) fragments are overrepresented. The main difference is in halogen containing groups: although they are present among the enriched fragments (sid.161, lnk.151), their frequency is much lower than among the ligands obtained using mol. weight as a controlling factor. Also, the total frequency of halogen containing compounds in the selected sets (17/66 and 14/69) is not significantly different from the reference compounds sets obtained with heavy atoms (302/1620 and 374/1778), suggesting that the high enrichment of halogen containing fragments in our article was primarily the result of using mol. weight as the factor controlling ligand size.

REFERENCES

(1) Gómez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernández-Lobato, J. M.; Sánchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. ACS Cent. Sci. 2018, 4, 268.

(2) Mohammadi, S.; O’Dowd, B.; Paulitz-Erdmann, C.; Goertlitz, L. Penalized Variational Autoencoder for Molecular Design. ChemRxiv 2019, DOI: 10.26434/chemrxiv.7977131.v2.