PubChem and ChEMBL beyond Lipinski

Alice Capecchi, [a] Mahendra Awale, [a] Daniel Probst, [a] and Jean-Louis Reymond*[a]

Abstract: Seven million of the currently 94 million entries in the PubChem database break at least one of the four Lipinski constraints for oral bioavailability, 183,185 of which are also found in the ChEMBL database. These non-Lipinski PubChem (NLP) and ChEMBL (NLC) subsets are interesting because they contain new modalities that can display biological properties not accessible to small molecule drugs. Unfortunately, the current search tools in PubChem and ChEMBL are designed for small molecules and are not well suited to explore these subsets, which therefore remain poorly appreciated. Herein we report MXFP (macromolecule extended atom-pair fingerprint), a 217-D fingerprint tailored to analyze large molecules in terms of molecular shape and pharmacophores. We implement MXFP in two web-based applications, the first one to visualize NLP and NLC interactively using Faerun (http://faerun.gdb.tools/), the second one to perform MXFP nearest neighbor searches in NLP (http://similaritysearch.gdb.tools/). We show that these tools provide a meaningful insight into the diversity of large molecules in NLP and NLC. The interactive tools presented here are publicly available at http://gdb.unibe.ch and can be used freely to explore and better understand the diversity of non-Lipinski molecules in PubChem and ChEMBL.

Keywords: chemical space, visualization, fingerprints, databases, biomolecules.

1 Introduction

PubChem and ChEMBL are public repositories of molecules and their biological activity.[1] While both databases contain a vast majority of small molecules, they also contain a small percentage of larger biomolecules such as peptides, oligonucleotides, oligosaccharides, and large natural products, as well as synthetic macromolecules such as peptide nucleic acids, fullerene derivatives, modified porphyrins and dendrimers. Such large molecules are interesting because they might serve as new modalities to address drug design problems which cannot be solved by small molecule drugs, for example blocking protein-protein interaction sites or delivering siRNA cargos into cells.[2] Unfortunately, PubChem and ChEMBL do not offer many options to explore these larger molecules. For instance, no overview of the database contents is provided, and the similarity search tools currently available on the respective websites focus on substructures, which is not well suited when relatively large molecules such as peptides are used as queries. An overview and searching across the entire content of this diverse family of large molecules is also not possible through specialized databases of biomolecules,[3] such as those for peptides,[4] oligonucleotides,[5] lipids,[6] or glycans.[7] Furthermore, the descriptions of chemical spaces for large molecules to date have remained focused on specific classes such as peptides and peptide macrocycles.[8]

Here we address this problem by designing web-based interactive tools to explore large molecules in PubChem and ChEMBL. We focus on molecules breaking at least one of the four Lipinski constraints for oral bioavailability (rule of 5: Molecular weight MW ≤ 500, number of hydrogen bond donor atoms HBD ≤ 5, number of hydrogen bond acceptor atoms HBA ≤ 10, calculated octanol/water partition coefficient clogP ≤ 5).[9] Although many orally available drugs, including peptides in particular, largely exceed Lipinski's rule of 5 limits,[10] Lipinski's criteria represent a useful definition to identify molecules that are clearly different from classical small molecule drugs. This concerns seven million of the 94 million entries in PubChem and 180,185 of the nearly 2 million entries in ChEMBL 24.1, which are described herein as the non-Lipinski PubChem (NLP) and non-Lipinski ChEMBL (NLC).[11]

To describe NLP and NLC, we aimed to create an interactive map of the databases and a similarity search tool to identify analogs of user-defined query molecules. We have previously reported interactive 2D- and 3D-maps and similarity search tools for a variety of small molecule databases.[12] In these applications, composition fingerprints such as MQN (Molecular Quantum Numbers)[13] and SMIfp (SMILES fingerprint)[14] provided readily interpretable maps when projected by principal component analysis (PCA).[15] Composition fingerprints also provide interesting associations between molecules in similarity search tools.[16] However, maps and similarity searches based on these composition fingerprints are not well suited for larger molecules. For example, they do not distinguish between peptides of different sequences if they have the same amino acid composition.

To obtain a meaningful classification of NLP and NLC, we use the principle of atom-pair fingerprints, which consider pairs of atoms and the distance separating them as structural features, and assign these features to bit values either by hashing or by counting.[12b, 17] Atom pair fingerprints tailored for small molecules such as CATS,[17c] Xfp[12b] and 3DXfp[18] have been shown to represent molecular shape and pharmacophores. Furthermore, we have already used atom pair fingerprints successfully to describe large molecules, in one case for detailed comparisons of 3D-models of

[a] *Department of Chemistry and Biochemistry, University of Bern*
Freiestrasse 3, 3012 Bern, Switzerland

**e-mail: jean-louis.reymond@dcb.unibe.ch*

biomacromolecules such as proteins and nucleic acids in the Protein DataBank (PDB),[19] and in the second case to perform virtual screening in libraries of peptide dendrimers and bicyclic peptides.[20] In the latter case our atom-pair fingerprint analysis perceives meaningful differences between peptides of identical composition but different sequences.

Here we introduce a new fingerprint denoted MXFP (macromolecule extended atom-pair fingerprint), which counts atom pairs in seven different categories up to topological distances exceeding 300 bonds. We show that MXFP is well suited to describe NLP and NLC in the form of two web-based applications. First, we present interactive chemical space maps based on MXFP featuring an easily interpretable classification of NLP and NLC. Second, we report a similarity search tool identifying analogs of any query molecules based on MXFP similarity. These web-based tools offer an unprecedented insight into the contents of NLP and NLC and reveal associations between large molecules which are otherwise difficult to identify.

2 Methods

2.1 Non-Lipinski subsets

Compound ID (CID) and SMILES were extracted for each entry in the PubChem Compound Database (downloaded April 5, 2018). For each entry, if more than one molecule was present, only the biggest fragment SMILES (based on its length) was considered for property and fingerprint calculations, however the entire SMILES was preserved. The SMILES were protonated (pH 7.4) using ChemAxon MajorMicrospecesPlugin (https://chemaxon.com). Hydrogen bond donor and acceptor count, cLogP and MW were computed for the largest fragment in each entry using RDKit, with *Lipinski*, *Descriptors* and *Crippen* modules respectively. All molecules violating more than one Lipinski rule were then classified as non-Lipinski. This led to 7,132,623 entries forming NLP and 183,185 entries forming NLC.

2.2 Property calculation

For each NLP and NLC entry atoms were classified into the following categories: heavy atom (HA), hydrophobic (HY), aromatic (AR), hydrogen bond acceptor and donor (HBA, HBD), positively and negatively charged (POS, NEG). AR, and HBA/HBD were assigned with, respectively, the ChemAxon TopologyAnalyzerPlugin, and the ChemAxon HBDAPlugin. HY was assigned to aromatic carbons, halogens, sulfur atoms without heteroatom neighbors, and to carbon atoms with at least one hydrogen atom neighbor. POS and NEG were assigned based on the atom formal charge.

2.3 Fingerprint calculation

MXFP is a 217D atom pair topological distance fingerprint calculated using an in-house Java program in a similar manner to our previously reported atom pair fingerprints 3DP and 2DP tailored for peptides and proteins.[19-20] Topological distances are measured using the TopologyAnalyzerPlugin provided as part of the JChem library by ChemAxon. There are seven atom categories: heavy atom (HA), hydrophobic (HY), aromatic atoms (AR), hydrogen bond acceptor and donor (HBA, HBD), positively charged (POS) and negatively charged (NEG), and only same-category atom pairs are considered. Each of the 217 values is the sum of contributions of atom pairs at a given distance for a given atom category. For each category *C*, all possible atom pairs *jk* contribute the value $g_{jk}(d_i)/s_{jk}$ to each of the 31 distance bins value v_{Ci} as described in Equation 1.

$$g_{jk}(d_i) = e^{-\frac{1}{2}\cdot\left(\frac{d_i - d_{jk}}{d_{jk}\cdot 0.09}\right)^2} \qquad s_{jk} = \sum_{i=0}^{30} g_{jk}(d_i) \qquad v_{Ci} = \frac{100}{{N_C}^{1.5}} \sum_{j=1} \sum_{k=1} \frac{g_{jk}(d_i)}{s_{jk}}$$

v = MXFP bin value
C = category
$C \epsilon \{HA, HY, AR, HBA, HBD, POS, NEG\}$
$i = \{i | i \epsilon \mathbb{N} \wedge 0 \leq i \leq 30\}$
N_C = total number of atoms in category C
$d_i = \{i | 0 \leq i \leq 6\}$
$d_i = \{d_{i-1} \cdot 1.18 | 7 \leq i \leq 30\}$
d_{jk}: topological distance between atoms j and k (1)

Atom pair distances d_{jk} are topological distances counted in bonds through the shortest path between two atoms. For each atom pair *jk*, $g_{jk}(d_i)$ is the value at distance d_i of a Gaussian of 18 % width centered on d_{jk} (Figure 2a). Gaussian values $g_{jk}(d_i)$ are sampled at the following 31 distance values d_i: 0, 1, 2, 3, 4, 5, 6, 7.1, 8.4, 9.9, 11.6, 13.7, 16.2, 19.1, 22.6, 26.6, 31.4, 37.1, 43.7, 51.6, 60.9, 71.8, 84.8, 100.0, 118.0, 139.3, 164.4, 193.9, 228.9, 270.0, 318.7. Each of these 31 gaussian values is normalized to the sum of all 31 values, s_{jk}, so that each atom pair contributes equally to the fingerprint. The sum of normalized gaussian contributions from all atom pairs of a certain atom category at distance d_i, is normalized by the number of category atoms to the power 1.5 to reduce the sensitivity of the fingerprint to molecule size, multiplied by 100 and rounded to unity to give the final fingerprint bit value v_{ci}. The 31 fingerprint bit values from each of the 7 atom categories are finally corrected by a category specific factor and joined, yielding the 217D fingerprint vector. In this work, we corrected the fingerprint bit values for the heavy atoms (HA) and aromatic atom (AR) categories by a factor 0.5 because the bit values were too high relative to the other atom categories. We calculated MXFP for the largest fragment in each NLP entry, but retained the complete SMILES in each entry.

2.4 Linearity calculation

The linearity of molecule *m*, *L(m)*, is a descriptor derived from MXFP. *L(m)* is defined as the ratio of *w(m)* and *w(a)*, where *a* is the linear alkane with the same number of heavy atoms as *m*, and *w* is the weighted mean of MXFP HA category, calculated according to equation (2).

$$w = \frac{\sum_{i=0}^{30}(i+1)\cdot v_{HAi}}{\sum_{i=0}^{30} v_{HAi}} \qquad L(m) = \frac{w(m)}{w(a)}$$

w = weighted mean of MXFP HA category
$i = \{i | i \epsilon \mathbb{N} \wedge 0 \leq i \leq 30\}$
$v_{HA} = MXFP$ bin value in HA category
m = analysed molecule
a = linear alkane with m HAC (2)

2.5 Similarity map calculation

Reference molecules were selected by sampling NLP across value triplets (HAC, AR/HAC, linearity) covering the range of

each of these three descriptors in 10% value increments. One compound was selected at random in each of the 1,000 resulting value triplets, which provided 324 reference molecules (676 of the value triplets did not contain any entry). For each database entry, the city-block distance in the MXFP chemical space, CBD_{MXFP}, to each of these 324 reference molecules was then calculated, giving a 324D NLP similarity space. The same approach was used to select reference molecules for NLC. Of the 1,000 triplets, 800 were discarded as they were not occupied by any entry, leading to a 200D NLC similarity space. The 324 NLP and 200 NLC references have very diverse structures (SMILES are provided in the SI).

2.6 Visualization in Faerun

The first three PCA components of the NLP and NLC similarity spaces were visualized in Faerun (variance covered, respectively: PC1 49%, PC2 28%, PC3 8%, and PC1 70%, PC2 15%, PC3 6%). A plain text file containing CID, SMILES, fingerprint, and properties of each NLP entry was processed using the Faerun preprocessing chain, which also includes a PCA service. Then Faerun was run using a docker container (https://github.com/reymond-group/faerun). Color coding of the Fearun map was enabled for HAC, HY/HAC (hydrophobic atoms fraction), AR/HAC (aromatic atom fraction), HBA/HAC (H-bond acceptor fraction), HBD/HAC (H-bond donor fraction), POS/HAC (positive charged atoms fraction), NEG/HAC (negative charged atoms), C/HAC (carbon fraction), RBC (rotatable bond count), CY/HAC (cyclic atom fraction), MW, HBA, HBD and clogP. Fraction values of atom categories are calculated from MXFP values, carbon fraction, rotatable bonds, cyclic fraction, MW, HBA, HBD and clogP are calculated with RDKit.

2.7 NLP-NLC comparison

20,000 entries were randomly picked from NLP or NLC and cut in two subsets of 10,000 entries each, A and B. Five series of 10,000 CBD_{MXFP} distances were then calculated as follows: a) A_{NLP} to the entire NLC, keeping the smallest non-zero value in each case; b) A_{NLC} to the entire NLC, keeping the smallest non-zero value in each case; c) A_{NLP} to B_{NLP}; d) A_{NLC} to B_{NLC}; e) A_{NLP} to A_{NLC}.

2.8 Similarity Search

The similarity search tool is a Python Flask (http://flask.pocoo.org/) app which uses Annoy (https://github.com/spotify/annoy) to search the MXFP NLP and NLC chemical spaces. Annoy is a C++ library with Python bindings developed by Erik Bernhardsson. Given its high speed and low memory requirements, Annoy was used to create two separate Annoy search files of NLP and NLC (for both, using *n_trees* = 50, *matrix* = Manhattan). In each similarity search instance, the user chooses to search NLP or NLC, and the correspondent Annoy file is selected. The Annoy file is used by the web app (with *search_k* = default) to retrieve the compound IDs of a pool of nearest neighbors (the no. of molecules to retrieve is a user choice). Then the compound IDs are associated back to the correspondent PubChem or ChEMBL SMILES. The results are displayed using SmilesDrawer.[21] The Similarity Search code is available open source at https://github.com/reymond-group/SimilaritySearch.

3 Results and Discussion

3.1 Non-Lipinski subsets

We define non-Lipinski molecules as those breaking at least one of the four Lipinski criteria (MW ≤ 500, HBD ≤ 5, HBA ≤ 10, clogP ≤ 5). For each PubChem and ChEMBL entry we applied the analysis to the largest molecular fragment, ignoring counter ions in the case of salts. When applied to the currently 94 million PubChem entries, these criteria selected 7,132,623 entries, which are defined here as NLP. NLP is a diverse set, with MW spanning from 181.15 Da to 19511.8 Da, clogP from -219.4 to +132.4, HBA from 0 to 442, and HBD from 0 to 235 (Figure 1, in green). The same analysis applied to ChEMBL led to 183,185 molecules, defined here as NLC, with MW spanning from 298.1 to 10173.49 Da, clogP from -67.9 to +101.8, HBA from 0 to 286, and HBD from 0 to 124 (Figure 1, in magenta).

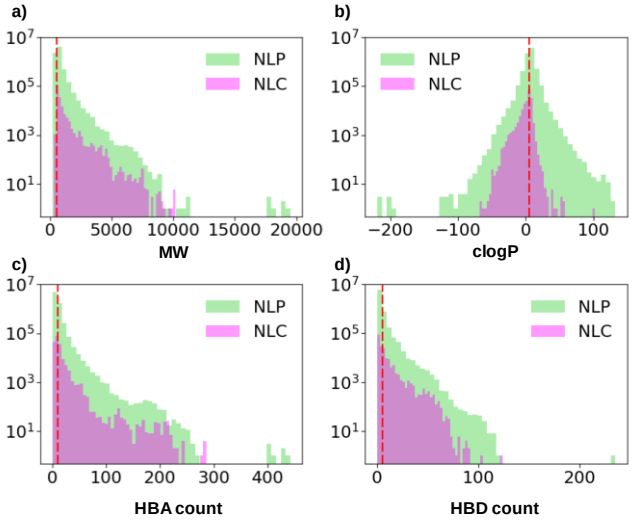

Figure 1. 1D-Histrograms of NLP (green) and NLC (magenta). a) MW, b) clogP, c) HBA, d) HBD. the vertical red dashed line indicates Lipinski's rule thresholds (MW = 500 Da, clogP = 5, HBA = 10, HBD = 5).

3.2 Macromolecule extended atom-pair fingerprint MXFP

MXFP is a 217D fingerprint counting atom-pairs using a fuzzy approach to assign atom-pairs to distance bins as done previously in our analysis of proteins and peptides.[19-20] In the case of proteins, we used an atom-pair fingerprint called 3DP which considers through-space distances between atoms in experimental 3D-structures from the Protein Databank.[19] To analyze peptides, we adapted our approach to use topological distances between residues in a related fingerprint called 2DP, with all atoms in a residue positioned at the α-carbon atom.[20a] For both fingerprints we consider four categories of atom pairs deemed essential for peptides, namely all heavy atoms (HA), hydrophobic (HY), positively charged (POS), and negatively charged atoms (NEG). For the MXFP presented here, we compute exact topological distances between atoms, which is suitable for any molecule. Furthermore, we use seven atom categories by additionally computing aromatic (AR), H-bond

donor (HBD), and H-bond acceptor atoms (HBA), which are important to differentiate molecules such as polycyclic aromatic hydrocarbons, oligosaccharides, and oligonucleotides. As for 3DP and 2DP, we do not consider cross-category atom pairs in MXFP.

MXFP values are calculated using the same approach and the same parameters as used previously for 3DP and 2DP. In detail, each atom pair is converted to a Gaussian of 18 % width centered at the atom pair topological distance, which is the shortest path between the two atoms counted in bonds. This Gaussian is then sampled at 31 distances d_i spanning from d_0 = 0 to d_{30} = 317.8 bonds at exponentially increasing intervals (Figure 2a). The sampled Gaussian values are normalized and added to the MXFP distance bins for the corresponding atom-pair category, and distance bins of each category are normalized to size (see Methods and Equation 1 for details). Sampling atom-pair Gaussians at exponentially increasing distances allows to describe molecules up to a very large size using only a limited number of dimensions in the atom pair fingerprint. The approach furthermore partly erases differences between atom pairs separated by a similar number of bonds at large distances, which favors the perception of global molecular shape over structural detail.

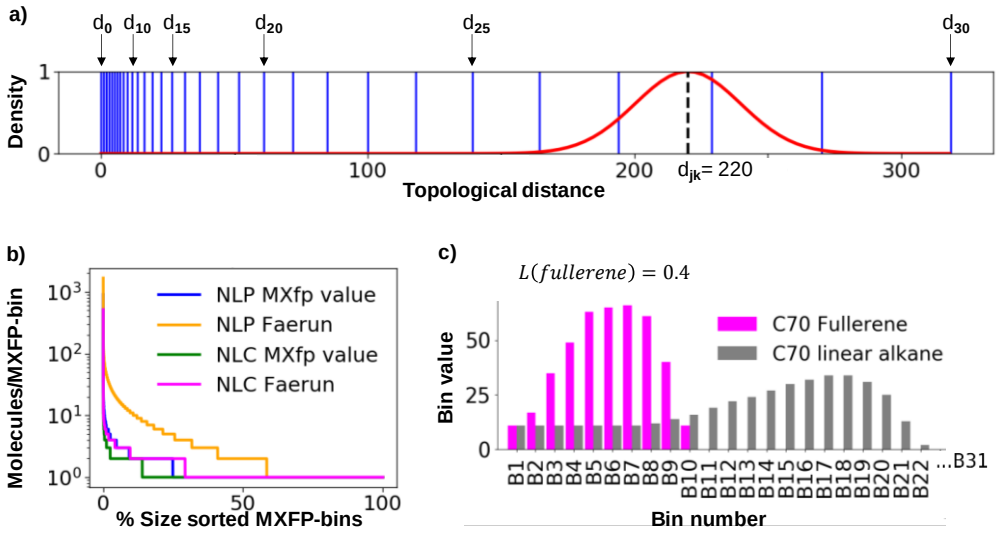

Figure 2. (a) In red Gaussian g_{jk} for an atom pair at topological distance d_{jk} = 220. In blue the 31 distances d_0 to d_{30} at which g_{jk} is sampled for calculating the contributions to MXFP. (b) Distribution of database entries in the unique MXFP-value bins (NLP in blue, NLC in green) and in the Faerun bins (NLP in orange, NLC in magenta). (c) Values of the first 31 MXFP bits for C70 Fullerene (magenta) and C70 linear alkane (grey).

The 7,132,623 molecules in NLP correspond to 4,753,197 unique MXFP value bins. The occupancy of the MXFP bins follows a power law distribution with 75% of the bins containing only a single NLP entry (blue line, Figure 2b). A similar molecules/MXFP-bins distribution is found for NLC, where the 183,185 molecules correspond to 153,616 unique MXFP values bins (green line, Figure 2b). The multiply occupied MXFP bins mostly contain entries sharing the same largest molecular fragment, or molecules with different structures but identical MXFP values such as diastereomeric carbohydrates (MXFP does not consider stereochemistry), or molecules with identical frameworks but different degrees of unsaturation such as lipids with fatty acids of equal length but different numbers of double bonds (MXFP does not distinguish non-aromatic carbon atoms with different degree of unsaturation). Note that grouping salts of the same compound with different counter ions, diastereoisomers of the same molecule or molecules only differing in the number of non-aromatic double bonds, makes perfect sense in the perspective of an analysis aiming at providing an overview of the database rather than a unique identifier for each entry.

Atom-pair fingerprints such as MXFP perceive molecular shape because spherical or cyclic molecules have a larger number of atom-pairs separated by short distances compared to linear molecules of the same size. Here we define a linearity descriptor L as a measure of topological molecular shape derived from the MXFP fingerprint. The linearity *L(m)* of molecule *m* is defined as the ratio of the weighted mean of the heavy atom pair bin index in the MXFP of molecule *m*, *w(m)*, to the same value for a linear alkane a with the same number of heavy atoms, *w(a)* (equation 2). The linearity value is 1 for the linear alkane, and lower for more globular molecules, e.g. *L*(fullerene) = 0.4 (Figure 2c). The linearity does not depend on building a 3D-model of the molecule as for the principal moments of inertia,[22] and is applicable to any molecule independent of its conformational flexibility.

3.3 MXFP chemical space visualization in Faerun

To lower the dimensionality of MXFP for visualizing NLP and NLC, we first attempted a direct principal component analysis (PCA) of the two datasets, however the first three PCs only gave partial coverage of data variance (48 % and 49%, respectively). We therefore constructed a representation based on the principle of similarity mapping.[23] Similarity mapping involves calculating similarities to a series of reference molecules in order to create a high-dimensional similarity fingerprint, which is then projected to lower dimensions by principal component analysis (PCA). The approach is interesting because the calculation of similarity maps is much faster than other dimensionality reduction methods for visualizing chemical space,[24] and is therefore applicable to very large datasets. Furthermore, many high-dimensional fingerprints, including MXFP, do not project well

into lower dimensions if PCA is applied directly to the fingerprint values, even when adding molecules with extreme properties, called satellites, as introduced by Oprea and coworkers.[25] However the projection of the corresponding similarity space often produces good results.

Similarity maps calculated by randomly choosing a few hundred reference molecules usually provide an approximately constant representation which is independent of the choice of references. However, the representation can be optimized for specific purposes by selecting the reference molecules, for example series of active compounds to visualize a structure-activity relationship study,[26] or references obtained by sampling regularly across important molecular properties to produce an ordered overview of a dataset.[27] Here we constructed similarity maps by calculating MXFP similarities to reference molecules selected across the range of MW, aromaticity and linearity covered by the NLP and NLC datasets, and then performing PCA (see methods). The procedure gave 3D-similarity maps covering 85% (NLP) and 91% (NLC) of the data variance, which was judged as sufficient to provide a good overview of the databases.

The 3D-similarity maps were then imported into Faerun, an open-source application recently reported by our group for rendering 3D-data interactively on the web.[12g] In this application each molecule is represented as a sphere, color-coded by a selected property, while its molecular structure is displayed on hover using SmilesDrawer, a compact molecular drawing program.[21] These interactive 3D-maps enable rapid browsing through NLP and NLC to gain an overview of their contents. As for the MXFP space itself, the distribution of NLP and NLC entries into Faerun bins follows a power law (Figure 2b, NLP orange line, NLC magenta line). The high-resolution NLP map contains a total of 1,413,817 bins, corresponding to an average of five molecules per bin. Multiple occupancies per bin in the NLP similarity map occur in part for the same reasons as for the MXFP bins, but also due to rounding of coordinates in the similarity map since bins (spheres) are placed on a 500 × 500 × 500 grid. The similarity map of NLC contains 123,878 bins, with an average of 1 molecule per bin. Note that each bin (sphere) in the Faerun map can be opened in a separate tab showing the distribution of molecules in the similarity space at higher resolution.

The MXFP-similarity 3D-maps of NLP and NLC are best inspected by using the web-based view (http://faerun.gdb.tools/). We have color-coded these maps according to different descriptors from lowest (blue) to highest (magenta) value (see methods for details). A selection of images of these color-coded representations illustrate the organization of NLP (Figure 3) and NLC (Figure 4) in the MXFP similarity space.

NLP forms a curved 3D-shape resembling a wave in which the smallest molecules are grouped on one side of the wave's head, intermediate sized molecules occupy the rest of the wave's head and the wave's body, and the largest molecules form the wave's tail, as illustrated by color-coding according to molecule size (Figure 3a). Color-coding by aromatic atom fraction shows that the outer shell of the wave's head contains molecules with the highest fraction of aromatic atoms (magenta), which are mostly polycyclic hydrocarbons (Figure 3b). The same view shows that the inner shell along the entire wave contains molecules with very few aromatic carbon atoms (blue), which comprise many linear alkanes, polyethyleneglycols, polyamines, as well as peptides. The intermediate layer contains molecules with intermediate aromatic atom fraction values (green), which are linker-extended drug-type molecules in the wave's head containing the lower size range, and oligonucleotides at the edge of the wave's tail containing the largest molecules. Oligonucleotides at the wave's tail are well visible in the map, color-coded by the fraction of negatively charged atoms (Figure 3c). This map also shows a group of smaller and more compact anionic molecules within the wave's head, which correspond to a variety of aliphatic polyphosphates and polycarboxylates.

The NLP similarity map also separates molecules according to their shape as measured by the MXFP derived linearity descriptor *L* discussed above (Figure 3d). The narrow blue region at the wave's head corresponds to globular molecules with a high percentage of aromatic carbons such as fullerenes. A second narrowly defined region at the center of the inner shell is colored in magenta and features strictly linear molecules containing long alkyl or polyethylene-glycol chains without any branching points. Peptides and oligonucleotides appear at intermediate values of linearity (yellow), which reflects the fact that these molecules are multiply branched by the attachment of amino acid side-chains (peptides) and nucleosides (oligonucleotides) along the main peptide respectively phosphodiester chain.

The different compound families are nicely separated by color-coding by the fraction of carbon atoms (Figure 3e). The figure shows examples of polycyclic hydrocarbons (exabenzocoronene, carbon fraction = 1.0, magenta), carbohydrates (difucosyllacto-N-hexaose, carbon fraction = 0.56, blue), peptides (exenatide, carbon fraction = 0.62, green) and oligonucleotides (mipomersen sodium, carbon fraction = 0.50, blue). Note that a close inspection of the MXFP similarity map of NLP using Fearun reveals many entries that are obvious mistakes in the PubChem database. For example, most mipomersen structures in PubChem are not drawn as the correct phosphorothioates but as the incorrect phosphate thioesters. Further structures of doubtful identity are also visible that contain linear chains of nitrogen and oxygen atoms.

NLC forms a similar but more sparsely populated wave-shaped 3D-similarity map. As with NLP, molecular size increases when navigating the map from the wave's head to its tail (Figure 4a, HAC color code). Aromaticity is higher in the outer shell of the wave head and diminishes upon traversing

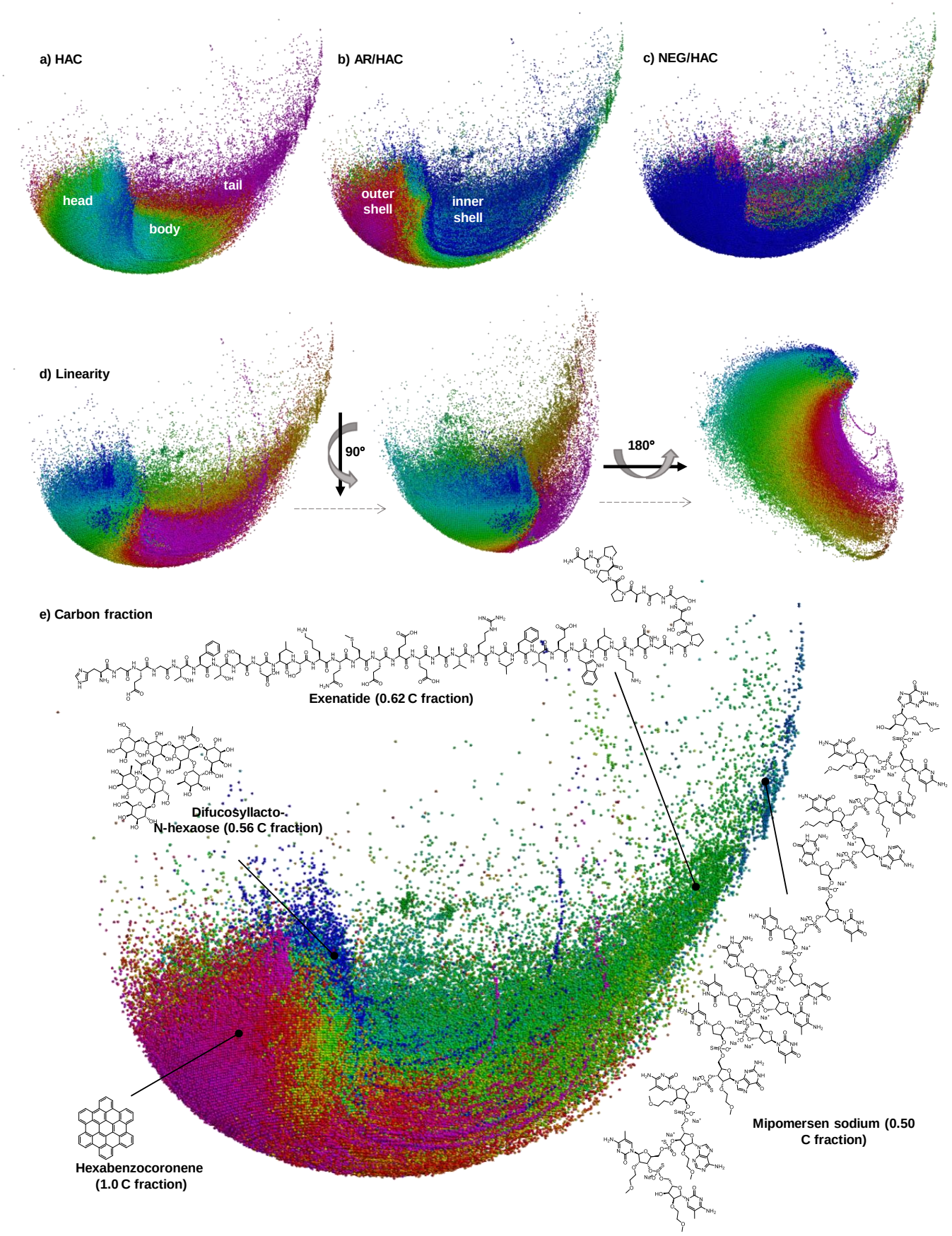

Figure 3. Similarity map of the NLP chemical Space colored using HAC (*a*), AR/HAC (*b*), NCHRG/HAC (*c*), linearity (*d*), and carbon fraction (*e*). In *d* are shown different rotation of the map. In *e* is shown the placement and the structure of hexabenzocoronene, difucosyllacto-N-hexaose, exenatide, and mipomersen sodium, as representative compounds of, respectively, polycyclic hydrocarbons, carbohydrates, peptides, and oligonucleotides.

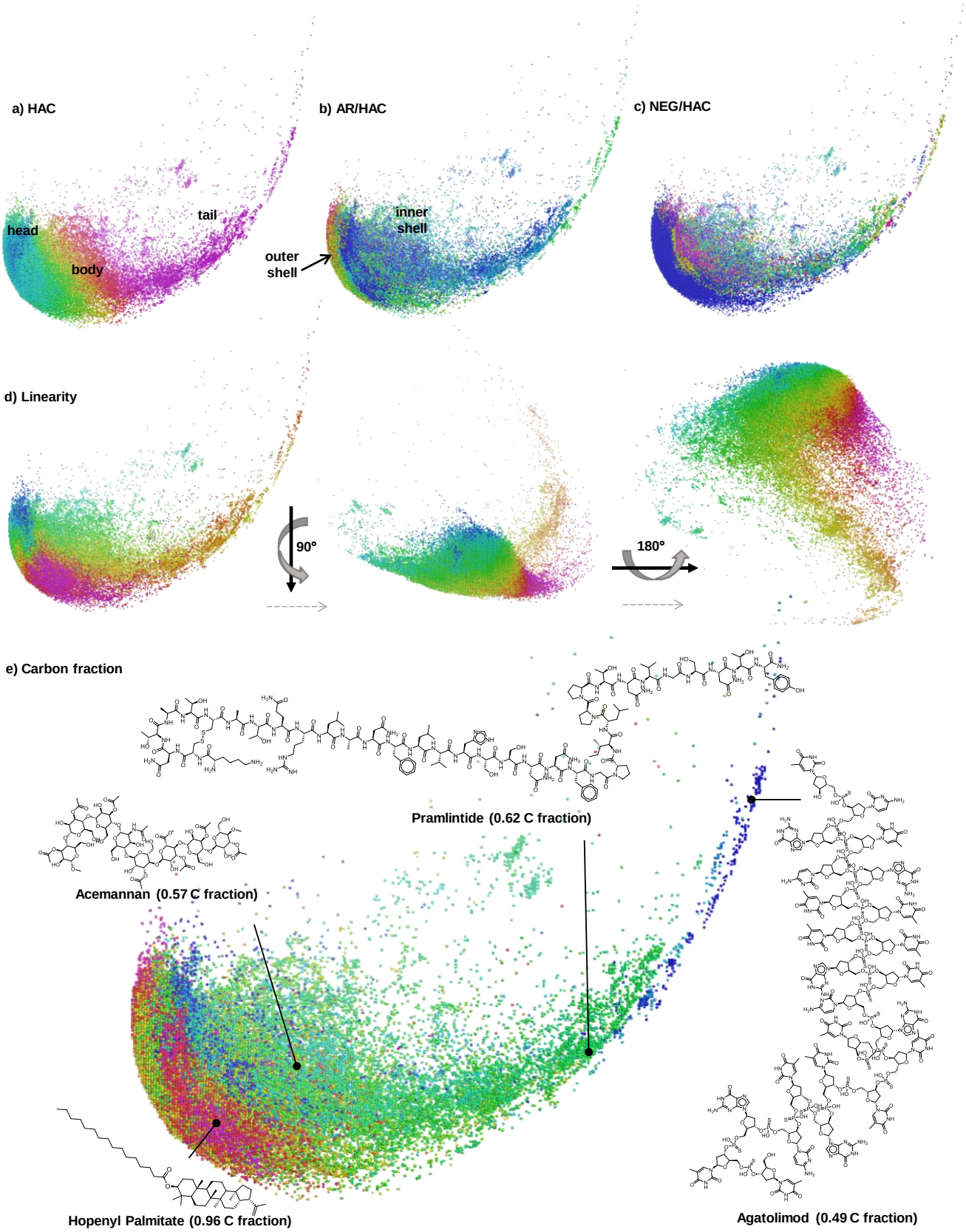

Figure 4. Similarity map on the MXFP NLC chemical Space colored using HAC (*a*), AR/HAC (*b*), NCHRG/HAC (*c*), linearity (*d*), and carbon fraction (*e*). In *d* are shown different rotation of the map. In *e* is shown the placement and the structure of hopenyl Palmitate, acemannan, pramlintide, andagatolimod, as representative compounds of, respectively, high carbon fraction molecules, carbohydrates, peptides, and oligonucleotides.

the map towards its inner shell (Figure 4b, AR/AHC color code). Compared to NLP (Figure 3b) the highly aromatic outer shell is less populated. Browsing this area in Fearun reveals an almost total absence of polycyclic hydrocarbons. As for NLP, the inner shell of the wave-shaped map contains mostly polyethylene-glycols, polyamines, and peptides, however the linear alkanes seen in NLP are mostly missing. As in NLP, the intermediate shell at the edge of the wave's tail with intermediate aromatic fraction (in green) contains oligonucleotides. Oligonucleotides are also well visible in blue using the NEG/HAC color code (Figure 4c). In terms of molecular shape, color coding by linearity *L* shows a similar distribution as for NLP (Figure 4d).

As with NLP, coloring the NLC similarity map by carbon fraction separates different compound families (Figure 3e). The figure shows examples of steroids (hopenyl palmitate, carbon fraction = 0.96, magenta), carbohydrates (acemannan, carbon fraction = 0.57, blue), peptides (pramlintide, carbon fraction = 0.62, green) and oligonucleotides (agatolimod, carbon fraction = 0.49, blue).

3.5 Comparing NLC with NLP

Because ChEMBL is one of the sources that feeds into PubChem, NLC represents a small (2.7%) subset of NLP.[28] To investigate if the remaining 97,8% of NLP cover a broader or different chemical space compared to this small NLC subset, we analyzed the CBD_{MXFP} distance distribution between 10,000 randomly picked NLP molecules and their NLC nearest neighbors, between NLC nearest neighbors, and between random pairs in NLP, NLC, and between NLP-NLC cross-pairs (Figure 5).

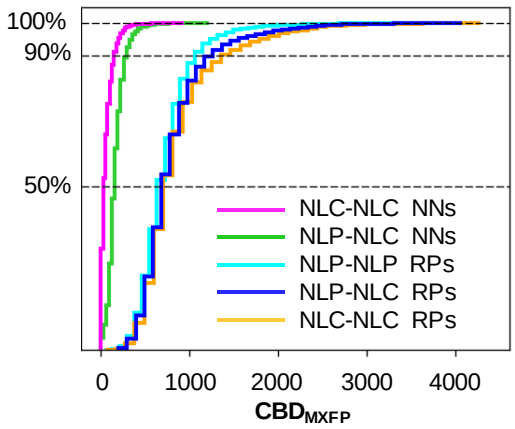

Figure 5. Distribution of CBD_{MXFP} distances between NLC and NLP molecules for nearest neighbors (NNs) and random pairs (RPs). See text and methods for details.

The analysis shows that 50% of NLP molecules have a nearest neighbor in NLC within CBD_{MXFP} 170, and more than 90% within CBD_{MXFP} 300 (Figure 5, green line), which is only a slightly larger distance distribution compared to the distance separating NLC nearest neighbors (Figure 5, magenta line). These nearest neighbor distances are much shorter than distances between random pairs of molecules within NLP (Figure 5, cyan line), within NLC (Figure 5, orange line), or between NLP and NLC molecules (Figure 5, blue line). We conclude that NLC and NLP cover a similarly broad chemical space, that NLC represents an almost random subset of NLP, and that NLP, although being 37-fold larger than NLC, does not cover a significantly different chemical space.

3.6 MXFP similarity search

PubChem currently offers a similarity search window in its beta version, which provides meaningful analogs of most query molecules. Unfortunately, this search option is designed for small molecules and fails to return any analog or does not return meaningful analogs when challenged with large molecules, most often when the query molecule is not itself present in PubChem. The same issue is experienced using the search function in ChEMBL. Examples of failed searches are shown in Table S1.

Here we designed an MXFP-similarity search tool for NLP and NLC as a web-portal using the approximate nearest neighbor search Annoy (Approximate Nearest Neighbors Oh Yeah, https://github.com/spotify/annoy) (Figure 6). This search option allows the user to browse NLP or its subset NLC and returns hundreds of MXFP-analogs of a query molecule in approximately 30 second per query. The similarity search tool is available at http://similaritysearch.gdb.tools/, and results are displayed on-screen using SmilesDrawer[21] and can be downloaded as a SMILES list.

The MXFP similarity search often returns results comparable to those provided by the PubChem and ChEMBL webpages whenever matched molecules have comparable substructures. However, compared to the PubChem and ChEMBL websites, which often fail to return results for unusual queries, MXFP similarity search provides a list of analogs in all cases. Analogs identified by MXFP similarity often comprise molecules with an overall molecular shape comparable to the query molecules, but with different structural composition. A good example is provided by searching NLP and NLC for analogs of **T7**, an antimicrobial peptide dendrimer with an unusual multi-branched peptide architecture which is active against multidrug resistant Gram-negative bacteria.[20c] While the PubChem and ChEMBL webpages do not find any meaningful analogs for this query, returning smaller and linear peptides, our MXFP similarity search points to related polycationic dendritic molecules of very different detailed structure. Besides many structures coming from patents, one example of interest at rank 4 in the search is CID 49775868, which is a peptide derivatized dendrimer of overall similar size, charge and shape as **T7**, but with very different detailed structure, reported to be active against HIV (Figure 6).[29]

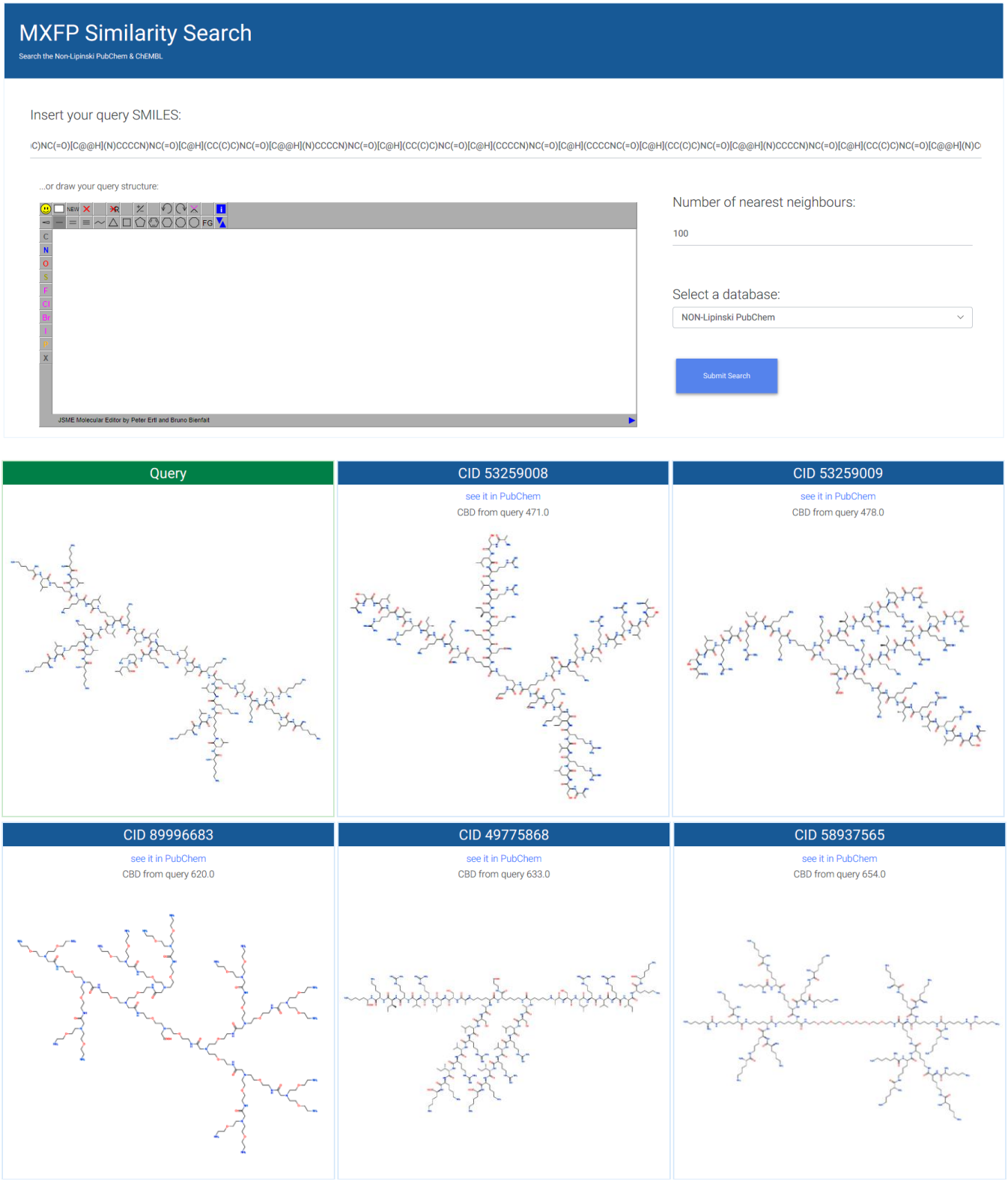

Figure 6. MXFP web interface. MXFP similarity search results for peptide dendrimer **T7**.

4 Conclusions

Here we focused on developing interactive tools to visualize and search large molecules in PubChem and ChEMBL breaking at least one of Lipinski's constraints for bioavailbility, defined here as NLP (7 million molecules) and NLC (180 185 molecules). We defined a 217D atom-pair fingerprint, MXFP, to describe these molecules in terms of molecular shape and pharmacophores. While MXFP is in principle suitable to describe molecules across the entire size range, here we focused on using this fingerprint to represent NLP and NLC in an interactive 3D-map and to enable a similarity search tool. These tools allow to rapidly browse through these diverse collections of macromolecules with unprecedented efficiency and identify interesting compound families and similarities between molecules which are otherwise difficult to perceive. The interactive tools presented here are publicly available at

http://gdb.unibe.ch and can be used freely to explore and better understand the diversity of non-Lipinski molecules in PubChem and ChEMBL.

Acknowledgements

This work was supported by a grant from the Vice-Rectorate Development of the University of Bern to A. C., and by the Swiss National Science Foundation. We thank ChemAxon Pvt. Ltd. for providing free academic and web licenses for their products.

References

[1] a) A. Gaulton, A. Hersey, A. Karlsson, D. Mendez, E. Cibrián-Uhalte, F. Atkinson, G. Papadatos, I. Smit, J. P. Overington, J. Chambers, L. J. Bellis, M. Davies, M. Nowotka, N. Dedman, P. Mutowo, A. R. Leach, A. P. Bento, M. P. Magariños, *Nucleic Acids Res.* **2016**, *45*, D945-D954; b) A. Gindulyte, B. A. Shoemaker, B. Yu, J. He, J. Zhang, J. Chen, L. Zaslavsky, P. A. Thiessen, Q. Li, S. He, S. Kim, T. Cheng, E. E. Bolton, *Nucleic Acids Res.* **2018**, *47*, D1102-D1109.

[2] H. Waldmann, E. Valeur, S. M. Gueret, H. Adihou, R. Gopalakrishnan, M. Lemurell, T. N. Grossmann, A. T. Plowright, *Angew. Chem., Int. Ed. Engl.* **2017**, doi: 10.1002/anie.201611914.

[3] D. J. Rigden, Xosé M. Fernández, *Nucleic Acids Res.* **2018**, *47*, D1-D7.

[4] a) T. Shtatland, D. Guettler, M. Kossodo, M. Pivovarov, R. Weissleder, *BMC Bioinformatics* **2007**, *8*, 280; b) J. Wang, X. Jiang, Y. Wang, T. Yin, X. Xiao, Z. Xue, D. He, *Database* **2018**, doi: 10.1093/database/bay1038.

[5] D. E. Newburger, G. Natsoulis, S. Grimes, J. M. Bell, R. W. Davis, S. Batzoglou, H. P. Ji, *Nucleic Acids Res.* **2012**, *40*, D1137-D1143.

[6] a) E. Fahy, S. Subramaniam, H. A. Brown, C. K. Glass, A. H. Merrill, Jr., R. C. Murphy, C. R. Raetz, D. W. Russell, Y. Seyama, W. Shaw, T. Shimizu, F. Spener, G. van Meer, M. S. VanNieuwenhze, S. H. White, J. L. Witztum, E. A. Dennis, *J. Lipid Res.* **2005**, *46*, 839-861; b) T. C. Kuo, Y. J. Tseng, *Bioinformatics* **2018**, *34*, 2982-2987.

[7] a) E. Gasteiger, F. Lisacek, J. Mariethoz, K. F. Aoki-Kinoshita, M. P. Campbell, R. Peterson, Y. Akune, N. H. Packer, *Nucleic Acids Res.* **2013**, *42*, D215-D221; b) J. Birch, M. R. Van Calsteren, S. Perez, B. Svensson, *Carbohydr. Polym.* **2019**, *205*, 565-570; c) O. Clerc, J. Mariethoz, A. Rivet, F. Lisacek, S. Perez, S. Ricard-Blum, *Glycobiology* **2019**, *29*, 36-44.

[8] a) W. M. Berhanu, M. A. Ibrahim, G. G. Pillai, A. A. Oliferenko, L. Khelashvili, F. Jabeen, B. Mirza, F. L. Ansari, I. ul-Haq, S. A. El-Feky, A. R. Katritzky, *Beilstein J. Org. Chem.* **2012**, *8*, 1146-1160; b) B. I. Díaz-Eufracio, O. Palomino-Hernández, R. A. Houghten, J. L. Medina-Franco, *Mol. Div.* **2018**, *22*, 259-267.

[9] C. A. Lipinski, F. Lombardo, B. W. Dominy, P. J. Feeney, *Adv. Drug Delivery Reviews* **1997**, *23*, 3-25.

[10] a) G. B. Santos, A. Ganesan, F. S. Emery, *ChemMedChem* **2016**, *11*, 2245-2251; b) B. C. Doak, B. Over, F. Giordanetto, J. Kihlberg, *Chem. Biol.* **2014**, *21*, 1115-1142; c) V. Poongavanam, B. C. Doak, J. Kihlberg, *Curr. Opin. Chem. Biol.* **2018**, *44*, 23-29; d) D. A. DeGoey, H.-J. Chen, P. B. Cox, M. D. Wendt, *J. Med. Chem.* **2018**, *61*, 2636-2651.

[11] a) Bradley C. Doak, B. Over, F. Giordanetto, J. Kihlberg, *Chem. & Biol.* **2014**, *21*, 1115-1142; b) P. D. Leeson, *Adv. Drug Delivery Reviews* **2016**, *101*, 22-33.

[12] a) M. Awale, R. van Deursen, J. L. Reymond, *J. Chem. Inf. Model.* **2013**, *53*, 509-518; b) M. Awale, J. L. Reymond, *Nucleic Acids Res.* **2014**, *42*, W234-W239; c) J. L. Reymond, *Acc. Chem. Res.* **2015**, *48*, 722-730; d) M. Awale, J. L. Reymond, *J. Cheminform.* **2016**, *8*, 25; e) M. Awale, J. L. Reymond, *J. Cheminform.* **2017**, *9*, 11; f) M. Awale, D. Probst, J. L. Reymond, *J. Chem. Inf. Model.* **2017**, *57*, 643-649; g) D. Probst, J. L. Reymond, *Bioinformatics* **2018**, *34*, 1433-1435.

[13] K. T. Nguyen, L. C. Blum, R. van Deursen, J.-L. Reymond, *ChemMedChem* **2009**, *4*, 1803-1805.

[14] J. Schwartz, M. Awale, J.-L. Reymond, *J. Chem. Inf. Model.* **2013**, *53*, 1979-1989.

[15] R. van Deursen, L. C. Blum, J. L. Reymond, *J. Chem. Inf. Model.* **2010**, *50*, 1924-1934.

[16] L. C. Blum, R. van Deursen, J. L. Reymond, *J. Comput.-Aided Mol. Des.* **2011**, *25*, 637-647.

[17] a) R. E. Carhart, D. H. Smith, R. Venkataraghavan, *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64-73; b) R. P. Sheridan, M. D. Miller, D. J. Underwood, S. K. Kearsley, *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 128-136; c) G. Schneider, W. Neidhart, T. Giller, G. Schmid, *Angew. Chem., Int. Ed. Engl.* **1999**, *38*, 2894-2896.

[18] M. Awale, X. Jin, J. L. Reymond, *J. Cheminf.* **2015**, *7*, 3.

[19] X. Jin, M. Awale, M. Zasso, D. Kostro, L. Patiny, J. L. Reymond, *BMC Bioinformatics* **2015**, *16*, 339.

[20] a) I. Di Bonaventura, X. Jin, R. Visini, D. Probst, S. Javor, B.-H. Gan, G. Michaud, A. Natalello, S. M. Doglia, T. Kohler, C. van Delden, A. Stocker, T. Darbre, J.-L. Reymond, *Chem. Sci.* **2017**, *8*, 6784-6798; b) I. Di Bonaventura, S. Baeriswyl, A. Capecchi, B. H. Gan, X. Jin, T. N. Siriwardena, R. He, T. Kohler, A. Pompilio, G. Di Bonaventura, C. van Delden, S. Javor, J. L. Reymond, *ChemComm* **2018**, *54*, 5130-5133; c) T. N. Siriwardena, A. Capecchi, B. H. Gan, X. Jin, R. He, D. Wei, L. Ma, T. Kohler, C. van Delden, S. Javor, J. L. Reymond, *Angew. Chem., Int. Ed. Engl.* **2018**, *57*, 8483-8487.

[21] D. Probst, J. L. Reymond, *J. Chem. Inf. Model.* **2018**, *58*, 1-7.

[22] W. H. Sauer, M. K. Schwarz, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 987-1003.

[23] a) A. S. Raghavendra, G. M. Maggiora, *J. Chem. Inf. Model.* **2007**, *47*, 1328-1240; b) J. L. Medina-Franco, G. M. Maggiora, M. A. Giulianotti, C. Pinilla, R. A. Houghten, *Chem. Biol. Drug. Des.* **2007**, *70*, 393-412; c) M. Awale, J. L. Reymond, *J Chem Inf Model* **2015**, *55*, 1509-1516; d) J. J. Naveja, J. L. Medina-Franco, *F1000Res* **2017**, *6*, Chem Inf Sci-1134.

[24] a) A. M. Wassermann, M. Wawer, J. Bajorath, *J. Med. Chem.* **2010**, *53*, 8209-8923; b) H. A. Gaspar, I. I. Baskin, G. Marcou, D. Horvath, A. Varnek, *J. Chem. Inf. Model.* **2014**, *55*, 84-94; c) T. Sander, J. Freyss, M. von Korff, C. Rufener, *J. Chem. Inf. Model.* **2015**, *55*, 460-473.

[25] a) T. I. Oprea, J. Gottfries, *J. Comb. Chem.* **2001**, *3*, 157-166; b) J. Rosen, J. Gottfries, S. Muresan, A. Backlund, T. I. Oprea, *J. Med. Chem.* **2009**, *52*, 1953-1962.

[26] C. Delalande, M. Awale, M. Rubin, D. Probst, L. C. Ozhathil, J. Gertsch, H. Abriel, J.-L. Reymond, *Eur. J. Med. Chem.* **2019**, *166*, 167-177.

[27] R. Visini, J. Arus-Pous, M. Awale, J. L. Reymond, *J. Chem. Inf. Model.* **2017**, *57*, 2707-2718.

[28] a) S. Kim, P. A. Thiessen, E. E. Bolton, J. Chen, G. Fu, A. Gindulyte, L. Han, J. He, S. He, B. A. Shoemaker, J. Wang, B. Yu, J. Zhang, S. H. Bryant, *Nucleic Acids Res.* **2016**, *44*, D1202-D1213; b) D. Yonchev, D. Dimova, D. Stumpfe, M. Vogt, J. Bajorath, *Drug Discov. Today* **2018**, *23*, 1183-1186.

[29] I. Bon, D. Lembo, M. Rusnati, A. Clo, S. Morini, A. Miserocchi, A. Bugatti, S. Grigolon, G. Musumeci, S. Landolfo, M. C. Re, D. Gibellini, *PLoS One* **2013**, *8*, e76482.