Occurrence of 'Natural Selection' in Successful Small Molecule Drug Discovery

A. Lina Heinzke[1], Axel Pahl[2], Barbara Zdrazil[1], Andrew R. Leach[1], Herbert Waldmann[3,4], Robert J. Young[5], Paul D. Leeson[6,*]

1. European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridgeshire CB10 1SD, United Kingdom
2. Compound Management and Screening Center, Max-Planck-Institute of Molecular Physiology, Otto-Hahn-Str. 11, 44227 Dortmund, Germany
3. Department of Chemical Biology, Max-Planck-Institute of Molecular Physiology, Otto-Hahn-Straße 11, 44227 Dortmund, Germany
4. Faculty of Chemistry and Chemical Biology, Technical University Dortmund, Otto-Hahn-Straße 6, 44227 Dortmund, Germany
5. Blue Burgundy Ltd, Ampthill, Bedfordshire MK45 2AD, United Kingdom
6. Paul Leeson Consulting Ltd, Nuneaton, Warwickshire CV13 6LZ, United Kingdom

Introduction

The historical application of molecules of natural origin as starting points for drug discovery[1] has largely been replaced by today's primary hit-seeking strategies of library screening (diversity-based, fragment-based, knowledge-based, virtual) and exploitation of known compounds.[2,3] Accordingly, the computed natural product (NP) probability score of approved oral drugs has fallen significantly for drugs invented after 1990,[4] when primary in vitro screening at cloned human targets began to be widely adopted. Despite this change, NPs, their derivatives, and synthetic compounds inspired by NPs make up 6%, 28% and 33% respectively of all small molecule drugs approved from 1981 to 2019.[5] This is consistent with the widespread appearance in approved drugs of NP partial structures and fragments.[6] Opportunities for the re-emergence of NPs in drug discovery have been widely heralded.[1, 4, 7-11]

Pseudo-natural products (PNPs) provide a new approach to quantifying the appearance of NP structural elements.[12-14] PNPs contain low molecular weight NP fragments, selected from a designed library,[15] which are connected in defined ways not currently known to be achievable by biosynthetic pathways. The PNP concept has been validated by their appearance in the literature[16,17] and by the design of several new classes of biologically active compounds.[18,19] Thousands of PNPs are available from commercial sources, and a PNP screening library can be readily established.[17]

Some 90% of approved drugs have a Tanimoto similarity of >0.5 to their structurally closest human endogenous metabolite, and screening collections were found to be less 'metabolite-like' than drugs.[20] NP-likeness has been proposed[21] to assist drug permeation via transporters, which evolved to facilitate entry of beneficial exogenous natural molecules, helping achieve selective tissue access and therapeutic efficacy. The aim of this work is to seek further support for the existence of 'natural selection' in drug discovery. We employ a highly curated dataset[22] from ChEMBL (version 32)[23] to assess the impact of quantitative NP measures, including the presence of PNPs, in marketed drugs and phase 1-3 clinical compounds in comparison with a background of relevant, target-matched reference compounds.

Methods

The general approach used is similar to our previous analysis of 'drug-like' properties,[24] but with the following notable differences in the dataset assembly:[22] 1) a newer version of ChEMBL was used (version 32), with the dataset limited to published literature information only (excluding a large amount of patent data associated predominantly with kinases which was added to ChEMBL in 2013-

16[22]); 2) clinical compounds in phases 1-3 as well as approved drugs (phase 4) are examined; 3) both clinical and background reference compound sets are carefully time-matched using first literature publication dates, with the emphasis here on recent practice (post-1990 and post-2008, see Results section); 4) the contemporary dataset is significantly larger, with >1000 clinical compounds and drugs published since 2008 versus the 141 drugs first approved in 2010-2020 used previously.[24]

Approved drugs (phase 4) and phase 1-3 clinical compounds, here collectively called Clinical compounds, were curated from ChEMBL version 32 as described.[22] In addition, non-clinical compounds were curated, here called Reference compounds, limited to those compounds with reported activity at the annotated biological targets understood to be responsible for the efficacy of the Clinical compounds.[22] Reference compounds qualified for entry only if they had a recorded pChEMBL value for in vitro activity at one or more of the Clinical compound's targets; Clinical compounds, already having published annotated targets, did not require pChEMBL values. Biological targets included mutated versions, and took into account the originating organism (95% were human targets); targets are defined in this paper by unique 'target name_mutant_organism' identifiers. Target classes were exhaustively identified[22] and further consolidated here to 17 major groups. Kinases and G-protein coupled receptors (GPCRs, subdivided into aminergic, peptidic and others based on their ligands) were the largest target classes, making up ~25% each of the post-2008 dataset, followed by transferases, nuclear receptors, proteases, oxidoreductases, and 8 smaller target classes.

Because changes to drug properties over time are significant,[24, 25] we aimed to ensure that Clinical and Reference compounds were strictly compared in time-matched periods. A Journal publication date was necessary for all entries; for Reference compounds, this was extracted from ChEMBL directly. For Clinical compounds, the dates of the first disclosure (normally the patent) were obtained from Scifinder® (CAS SciFinder^n - Chemical Compound Database | CAS) and used for analysis of long-term trends. The first Journal publication dates of Clinical compounds, where not in ChEMBL for post-1990 first disclosure compounds, were obtained from Scifinder®. The median Journal publication date for post-1990 Clinical compounds was five years after the first disclosure.

The full dataset, after removal of molecules filtered by the PNP algorithm[17] (predominantly molecular weight >1000 and presence of uncommon elements) contained 3173 unique Clinical compounds and 388,027 unique Reference compounds. In all there were 9644 Clinical compound-target pairs, 596,341 Reference compound-target pairs, covering 2285 targets. For the post-2008 Journal publication period, used for Clinical versus Reference compounds analysis (see Results section), there were 1212 unique Clinical compounds and 229,569 unique Reference compounds,

comprising 2842 Clinical compound-target pairs, 320,927 Reference compound-target pairs, and 726 targets.

For analysis by individual mean or median Reference compound properties by target, we required the target to have ≥ 100 Reference compounds.[22,24] The post-2008 set contained 422 targets with ≥ 100 Reference compounds, acted on by 1091 Clinical compounds unique to each target class, comprising 2011 Clinical compound-target pairs; 28 of the 1091 compounds are duplicated because they act at more than one target class. The range of Reference compound numbers acting at the Clinical targets was 100-7484 (median 408). Of the 1091 Clinical compounds, 737 had one biological target, 132 had two targets and 222 (122 acting at protein kinases) had three or more targets (range 3-17).

For quantitation of NP character, three complementary measures were used:

1. **PNP_Status.** Compounds were assigned to one of four categories according to their NP fragment combination graphs.[16] The NP library fragments used for this purpose are Murcko scaffolds[26] (the core structures containing all rings without substituents except for double bonds, n=1673) derived[16] from a representative set of 2000 NP fragment clusters.[15] Because of their ubiquitous appearances in NPs, the phenyl ring and glucose moieties were specifically excluded as fragments.[16] The phenyl ring however does appear in some fragments, combined with other ring systems. The categories are:
    - **NP** (natural product). Naturally occurring compounds with defined structures and fragment combinations.
    - **NPL** (NP-like). Fragment connections appear as found in NPs, but the structures are different from NPs, e.g. NP derivatives, or compounds with additional NP fragments.
    - **PNP** (pseudo-natural product). Two or more NP fragments linked by 0-3 atoms in defined ways not found biogenetically. Where NP fragment combinations in a molecule had both NP and PNP motifs, they were assigned to the PNP category.
    - **NonPNP**. All others.
2. **Frag_coverage_Murcko**. This measure has no dependency on the connectivity between fragments, and is equal to the number of heavy (non-H) atoms (HA) present in the NP fragments divided by the total number of HA that are present in the Murcko scaffold of each molecule.
3. **NP-likeness**. A Bayesian measure of similarity to the structural space covered by natural products, calculated by the method of Ertl.[27]

PNP_Status is a compound categorization, whereas Frag_coverage_Murcko and NP-likeness have quantitative values for all compounds. Some illustrative examples of marketed drugs that are classified PNP, NonPNP and NPL, and the fragments they contain, are shown in Figure 1.

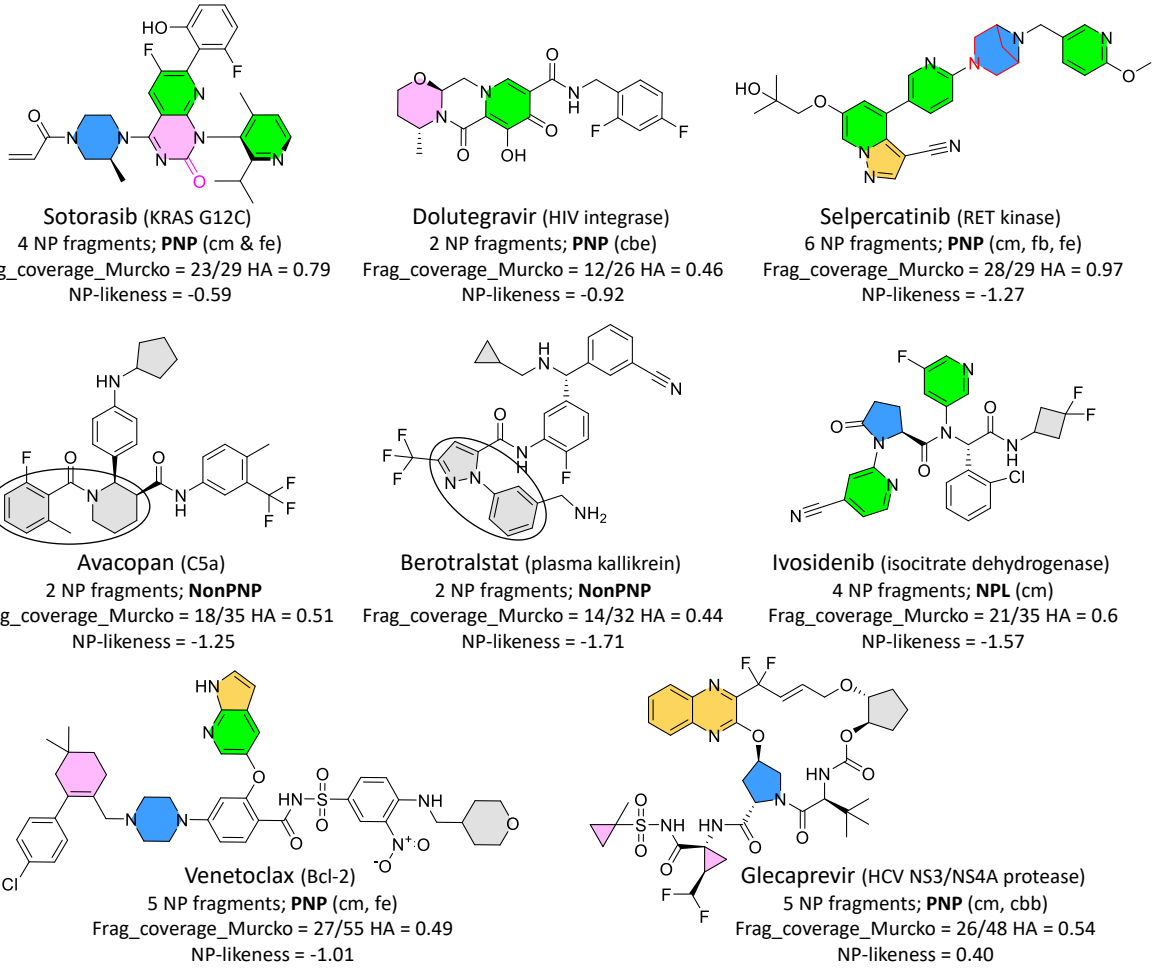

Figure 1. Examples of marketed drugs illustrating the NP metrics used. NP fragments are filled and those with >1 ring system are circled. Colour filled NP fragments show molecules classified as pseudo-natural products (PNP) or NP-like (NPL); see Methods section for details. In PNPs and NPLs, fragments are connected in several non-biogenetically accessible ways shown in parentheses (see ref 16 for complete definitions), using ≤ 3 atoms in PNPs, and biogenetically in NPLs. The most common PNP connectivities seen in the post-2008 Clinical compounds in this study are exemplified in the compounds shown: monopodal (single bonds, termed cm, 73.9%), ring fusion (termed fused edge, fe, 14.1%), bridged ring (termed fused bridge fb, 4.7%), and two bipodal options with one connecting ring (cbe, 2.2%; cbb, 1.4%). NonPNPs can contain NP fragments, connected by >3 atoms. Frag_coverage_Murcko values show the heavy or non-H atom counts (HA) in combined fragments and the Murcko scaffold.

In addition, the specific NP fragments, fragment combination pairs, and PNP fragment combinations that are found in Clinical and Reference compounds were identified in the post-2008 set, in full and by major target classes. Their relative Clinical versus Reference abundances were estimated using odds ratios. Compound physicochemical properties were added to the dataset,[22] taken from ChEMBL and RDkit.[28]

Calculations were performed using Microsoft Excel (https://www.microsoft.com/excel) and DataWarrior (www.openmolecules.org). Statistical significance (*p* values) was obtained from t-tests assuming unequal variance for unpaired data, or from Wilcoxon signed rank tests for paired data.

Source datasets: caveats and limitations

NP fragment library. This was generated for use in fragment based drug discovery (FBDD), with a focus on practical application and commercial fragment availability.[15] Although not originally intended to be used for statistical analysis of NP characteristics, it is used as such here because it forms the basis of the PNP classification algorithm.[16] The library was assembled starting from 183,769 NPs containing at least one ring, which contained 110,485 fragments.[15] Using pharmacophore, physical property and chemical alert filtering, this was reduced to 2000 representative clusters and further refined[16] to a set of Murcko scaffold fragments. The NP library is composed of 1673 fragments (MW 42-294) having either single ring systems (1421 fragments) or >1 ring system (252 fragments). Because of its targeted makeup, small size, and property distribution, the fragment NP library lacks NP acyclic substituents[29] and there is no directly comparable 'non-NP' fragment set that could be used as a control set. Representation of all NP chemical ring systems is an impractical proposition. For example, other similar studies found 134,102 fragments (MW 100-300) including acyclic moieties in a set of 210,213 NPs,[6] and 38,662 unique ring systems were found in 269,226 NPs, of which 23,299 (60%) were singletons.[30] High diversity is apparent in studies of scaffold occurrence: singleton scaffolds dominate the medicinal chemistry literature[31] and only 763 of 103,772 scaffolds (0.7%) in the ChEMBL 20 database were found in >10 compounds.[32] Not surprisingly, high proportions of singleton NP fragments, and singleton NP fragment combinations, are also seen in the Clinical and Reference compound sets (see Results). Of the 30 most abundant NP ring systems reported,[30] 21 passed the fragment filtering process applied[15,16] and are present in the NP fragment library.

CheMBL data. The freely available ChEMBL database abstracts comprehensive structure-activity data from the medicinal chemical literature and has become an established mainstay for chemoinformatic studies.[23] As discussed above, here we use only those Reference compounds that

are reported in scientific journals.[22] A major caveat[22, 24] is that compounds revealed in publications are generally selected to illustrate how various problems were addressed and solved, and therefore carry a risk of being unknowingly biased by author selection. A published medicinal chemistry case history is unlikely to be representative of all that was done, as in practice it discloses only a small proportion of all molecules synthesised in a project. Additionally, in describing structure-activity relationships, there may be a general tendency to emphasise more of the 'better' compounds (more potent, good pharmacokinetics) than the 'poorer' (less potent or metabolically unstable) compounds. While patented compounds are more likely to be representative of what was done, they often lack quantitative potency data, which we required for entry in the Reference compound set. Restricting the analysis to higher potency Reference compounds by introducing a pChEMBL cut-off could in principle improve quality. This was not done for two reasons: 1) it is not possible to compare relative potency values across very widely differing assay formats,[33] and 2) it would reduce the numbers of clinical compounds as well as clinical targets with ≥ 100 Reference compounds. Finally, the annotation of specific biological targets to Clinical compounds is based on current knowledge and could change in future.

Target class NP statistics. While trends in NP properties are clearly apparent (see Results), statistically significant differences ($p < 0.05$) are often absent for individual target classes with lower numbers of Clinical compounds and/or targets.

Impact of time on Clinical and Reference compound NP properties

The long-term progression of the fraction of Clinical compounds by phase reached in each of the PNP_Status categories (Figure 2a) shows near-identical time trends in all clinical phases (1-4), so Clinical compounds were combined into a single group for further analysis (Figures 2b-d). The consistent increase in PNP fraction is striking, approximately doubling every 2-3 decades, reaching 67% of all Clinical compounds in the 2010s (Figures 2a, b). Significant increases in Frag_coverage_Murcko values have occurred every second decade, with their distribution narrowing over time: after 2010 the average clinical compound has a value of 0.66, with 1st and 3rd quartile values of 0.50 and 0.84 respectively (Figure 2c). The fraction of NP compounds is falling over time (Figure 2b), which is consistent with the trend in NP-likeness, which falls significantly from historical levels in the 1960s-1990s, and further again from 2000 onwards (Figure 2d).[4] Interestingly, NPL compounds consistently appear in all time periods at ~10-20% of the total (Figure 2b). NonPNPs were the majority class until the 1990s; because PNPs and NonPNPs are the dominant classes overall, their fractional occurrence over time is inversely related. It is clear from Figure 2 that while application of NP structures *per se* has declined markedly, the 'NP signal' is nevertheless present,

expressed by the increasing application of NP-derived fragments, resulting in the current dominance of PNP Clinical compounds.

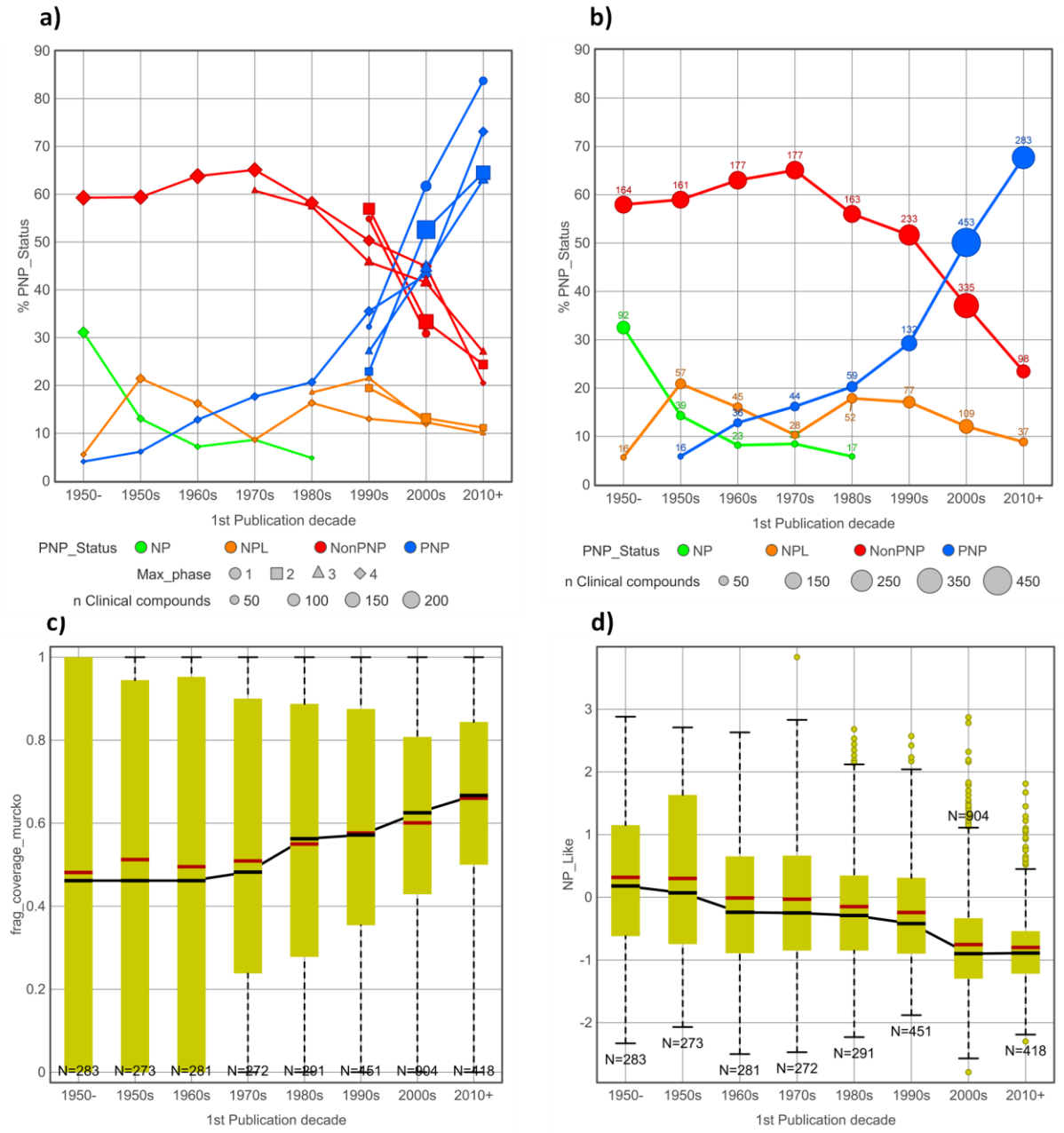

Figure 2. NP properties of Clinical compounds by decade of 1st publication (usually a patent). a) % Clinical compounds by Phase reached in each PNP_Status category. b) Same data as a) with Clinical phases combined. c) Frag_coverage_Murcko. Significant increases ($p < 0.05$, unpaired t-test) occur every 2nd decade from the 1970s. d) NP-likeness. Significant decreases ($p < 0.05$, unpaired t-test) occur in the periods 1960s-1990s and 2000s onwards.

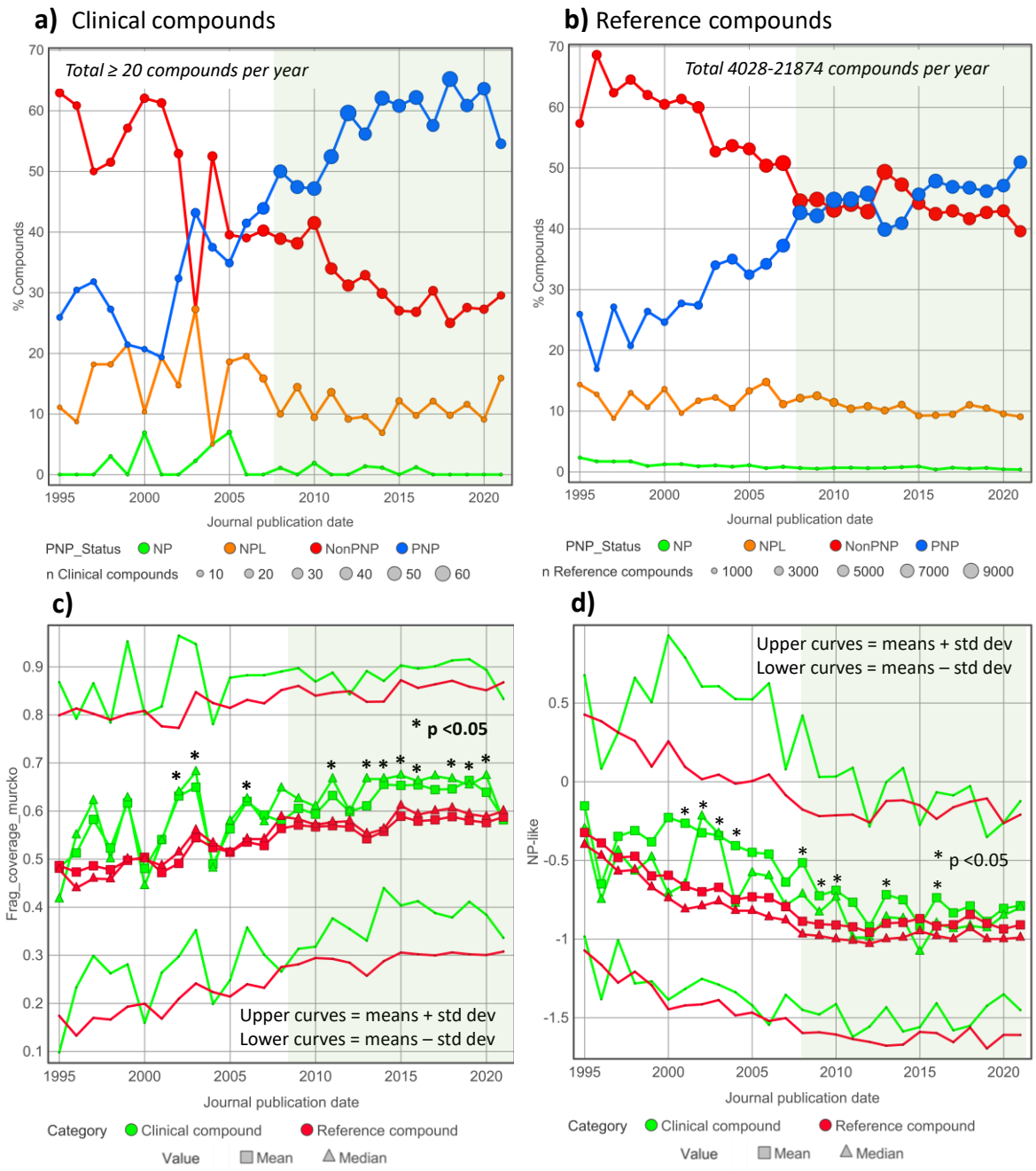

Figure 3. Clinical and Reference compound NP properties by first Journal publication date since 1990. Clinical and Reference compounds all act at the same targets. Clinical compounds all have first publication (usually patent) dates 1990 or later (the years 1990-4 have <20 Clinical compounds and are not shown). a) % Clinical compounds in each PNP_Status category. b) % Reference compounds in each PNP_Status category. c) Frag_coverage_Murcko. d) NP-likeness. *p* Values are from t-tests assuming unequal variance. Reference compound NP values do not alter post-2008 (shaded green), and this period was chosen for target class analysis.

The period from 1990 onwards was selected for comparing Clinical with Reference compounds by NP properties, because of both the sharp increase in numbers of Clinical PNPs (Figure 2b), and the availability of several thousand comparative ChEMBL Reference compounds each year. Figures 3a and 3b show the progression of the PNP_Status categories of the post-1990 Clinical compounds present in Figure 2b, and target-matched ChEMBL Reference compounds, both by first Journal publication date. While increases in PNP fraction occur in both Clinical (Figure 3a) and Reference (Figure 3b) compounds until ~2008, later Reference compounds show a markedly smaller increase than do Clinical compounds. Comparing Figures 3a and 3b shows a clear 'enrichment' of PNPs in Clinical compounds versus Reference compounds in the post-2008 period. It is also apparent from Figures 3a and 3b that the fractions of compounds in the NP and NPL categories are very similar in Clinical and Reference compounds and are not changing over time.

Consistent with the PNP time trend, Frag_coverage_Murcko (Figure 3c) values are increasing over time, with Clinical compounds having significantly higher values than Reference compounds in eight of the years after 2010. NP-likeness is decreasing over time, as expected, in both Clinical compounds than Reference compounds (Figure 3d). However, higher NP-likeness values for Clinical compounds versus Reference compounds are seen in nine of the years post-2000, although the differences are less pronounced in the post-2010 period. Comparing all Clinical and Reference compounds since 2008 confirms statistically significant increases in the three NP properties (Table 1). By odds ratio analysis of the full dataset, a post-2008 Clinical compound is 54% more likely to be a PNP than a Reference compound (Table 1).

Table 1. NP property values for Clinical and Reference compounds for the full post-2008 set.

a) % PNP and Clinical vs Reference Odds Ratio					
Category	**Clinical count (%)**	**Reference count (%)**	**Odds ratio**	**95% CI**	***p***
PNP	666 (57.3%)	101925 (46.6%)	1.54	1.37-1.73	<0.0001
Other	497 (42.7%)	116848 (53.4%)			
b) NP Properties: Clinical vs Reference Compounds					
Property	**Category**	**Mean**	**Median**	**Std. Dev.**	***p***
Frag_coverage_Murcko	Clinical	0.62	0.64	0.26	<0.0001
	Reference	0.58	0.59	0.28	
NP-likeness	Clinical	-0.77	-0.9	0.74	<0.0001
	Reference	-0.92	-1	0.71	

Collectively, Figures 3a-d indicate that Clinical compounds possess greater 'NP character', based on the three measures used, than do Reference compounds. Notably, for each of the NP metrics, there is relatively little change in their Reference compound annual values from 2008 onwards (shaded

green in Figures 3a-d). For this reason, the post-2008 period was selected for further analysis by target class.

Post-2008 target class NP properties

The results in Figure 2, 3 and Table 1 take no account of the biological targets followed, which are known to have significant impact on the physical properties of their ligands.[24,34] The role of each target class (n=17) on the NP profiles of post-2008 Clinical versus Reference compounds was examined in three ways:

1. Comparisons of all Clinical (n=1212) versus all Reference (n=229569) compounds. This takes no account of the widely differing numbers of reference compounds annotated to each target. As an example, in the kinase target class, this approach compares 330 Clinical compounds with 50370 Reference compounds.
2. Unpaired comparisons of all Clinical compounds (n=1091) versus the median values for their targets that possess ≥ 100 Reference compounds (n=422). This treats Reference compounds on an equal footing by target but does not completely account for Clinical compounds that have multiple targets. In the kinase target class example, 329 Clinical compounds are compared with the median values for 122 single kinase targets having ≥ 100 Reference compounds.[22,24]
3. Paired comparisons of all Clinical compounds versus the corresponding median values for their targets that possess ≥ 100 Reference compounds (n=1091).[24] While this results in certain target values being duplicated (especially kinases), if reflects the reality that some targets have been pursued more than others. For those Clinical compounds with >1 target, medians of the targets' median NP property values were used. In the kinase target class example, 329 Clinical compounds are compared with the 329 median values for their annotated targets having ≥ 100 Reference compounds.

It is apparent from the results of the paired analysis (i.e. 3 above), shown in Figure 4, that target class significantly influences all the NP properties and that, for the majority, Clinical compounds have higher NP property values than the corresponding Reference compounds. In addition, the rank orders of target class [Clinical-Reference] differences are different for each NP metric in Figure 4, suggesting that the three measures are complementary estimates of NP character.

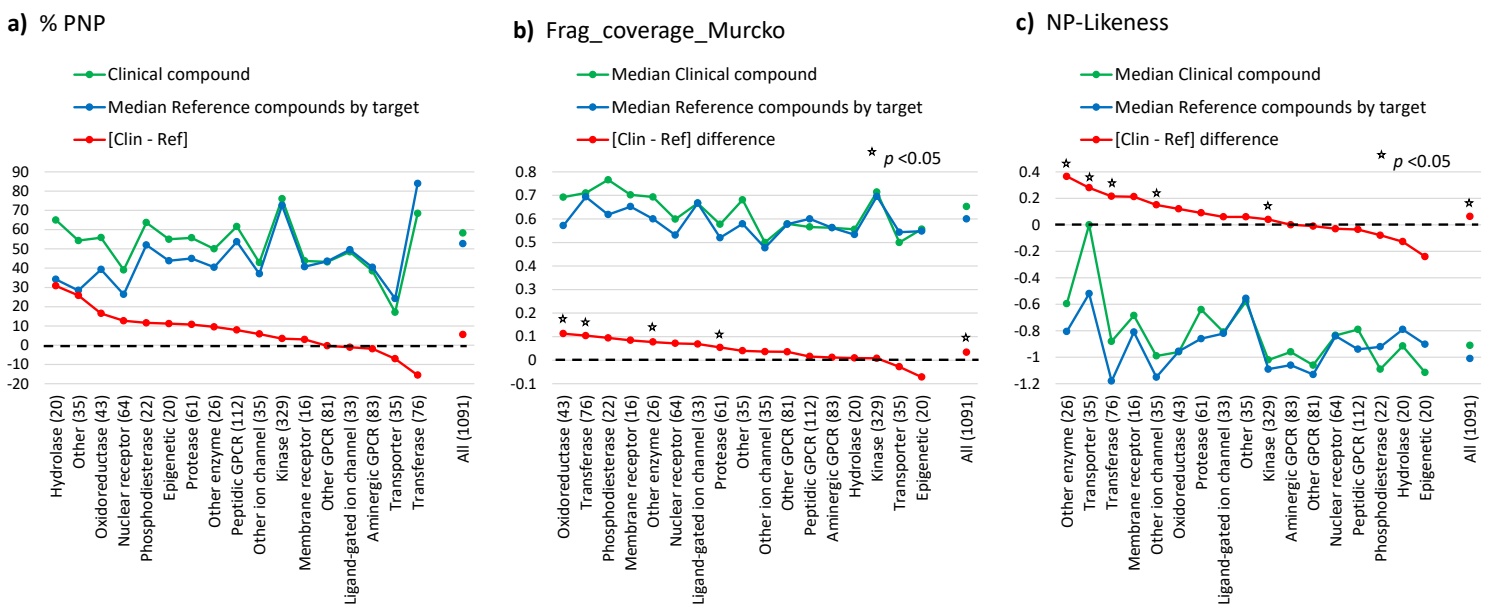

Figure 4. Post-2008 paired Clinical versus target Reference compound analysis of NP metrics, by target class. Clinical compounds in each target class are unique, and paired with their corresponding Reference compounds by target(s). Target values are median values for all targets having ≥100 Reference compounds. Target classes on the x-axis are shown with the numbers of clinical compound-target pairs. a) %PNP; this is categorical measure, and the Clinical-Reference differences (in red) are the arithmetic differences. b) Frag_coverage_Murcko. c) NP-likeness. For c) and d) the Clinical-Reference differences are the medians of the differences for each Clinical compound. The black dotted line is where Clinical and Reference values are equal. The rank orders of target class [Clinical-Reference] difference values differ according to the NP metric used. The correlation r^2 values (n=17) are: 0.003 for % PNP vs Frag_coverage_Mucko; 0.181 for % PNP vs NP-likeness; and 0.187 for Frag_coverage_Murcko vs NP-like. *p* Values were determined using the Wilcoxon signed rank test.

The corresponding analyses by all compounds and by unpaired targets (i.e. 1) and 2) above) are qualitatively very similar to Figure 4 (Figure S1). The collected % differences in the NP measures between Clinical and Reference compounds found by the three approaches (Table 2) show that increased NP properties in Clinical compounds versus Reference compounds consistently appear more frequently by target class than the opposite possibility. The three NP measures show differing results by target classes, for example epigenetic and transporter display opposite trends on % PNP and NP-likeness. In the case of Clinical transporter compounds, the NP-likeness values are increased by a group of 12 sodium/glucose cotransporter 2 (SGCL2) inhibitors (e.g. empagliflozin), which all possess a glucose part-structure (excluded as a NP fragment[16]). Across target classes, there is clear

variability in PNP content, with protein kinases and transferases being the most PNP-rich. Most kinase inhibitors bind competitively to the adenosine triphosphate (ATP) site, mimicking the adenosine heterocycle's hydrogen bond donor and/or acceptor interactions with the 'hinge' kinase domain. With transferases, the target diversity is rather limited as it is dominated by compounds acting at the four isoforms of phosphoinositol-3-kinase (PI3Kα, β γ and δ).

Table 2. Summary of Clinical vs Reference compound normalized NP property trends by target class.

Target class	% Differences between Clinical and Reference Compounds[a]								
	% PNP			Frag_coverage_murcko			NP-like		
	All cmpds	Tgt ≥100 unpaired	Tgt ≥100 paired	All cmpds	Tgt ≥100 unpaired	Tgt ≥100 paired	All cmpds	Tgt ≥100 unpaired	Tgt ≥100 paired
Other	30	80	91	18	32	5.3	15	29*	7.3
Hydrolase	107	90	90	16	5.9	2.0	4.4	9.0	-22
Nuclear receptor	28	88	48	22	34*	10	-9.2	0.6	-3.5
Oxidoreductase	73	52	42	19*	21	20*	4.9	-1.1	13
Epigenetic	45	76	25	0.2	6.2	-13	-16	-33	-25
Protease	69	100	24	20*	26*	11*	47*	11	11
Other enzyme	34	25	24	18*	21	13*	37*	26*	37*
Phosphodiesterase	3.8	-1.0	22	13	23	14	-5.8	-7.1	-8.0
Other ion channel	-31	23	16	-13	4.8	7.1	15*	12*	13*
Peptidic GPCR	14	15	15	-3.5	1.6	2.6	14	20	18
Membrane receptor	25	61	7.3	16	45	17	0.0	13	23
Kinase	13	2.4	4.7	3.1*	0.0	1.2	6.9*	2.9	4.3*
Other GPCR	-6.5	28	-0.8	-2.5	2.4	6.3	0.9	-9.3	-1.0
Ligand-gated ion channel	22	-2.1	-2.1	15*	1.5	10	12	0.0	7.3
Aminergic GPCR	-4.5	-2.0	-4.6	-1.6	2.0	1.5	7.1	-6.1	0.0
Transferase	-2.4	40	-18	10	11*	18*	29*	14	23*
Transporter	-42	-29	-29	-12	-7.1	-5.0	97*	100	10*
All	**21**	**21**	**10**	**9.2***	**9.3***	**5.6***	**10***	**6.2***	**6.8***

Clin >10% higher than Ref	Ref >10% higher than Clin
Clin >0 – 10% higher than Ref	Ref >0 – 10% higher than Clin

[a] Shown are the colour coded % differences between Clinical and Reference compound median NP properties, found using 1) all compounds, and Clinical compounds either 2) unpaired or 3) paired to targets containing ≥ 100 Reference compounds (see Results

section). The % values shown are equal to [(Clinical÷Reference)-1)]x100 for % PNP and Frag_coverage_Murcko, and because all values are negative, [-(Clinical÷Reference)+1)]x100 for NP-likeness. The Table is ranked by the paired % PNP score. The colour coding provides a qualitative guide to Clinical or Reference compound preference. * $p < 0.05$ (Wilcoxon signed rank tests).

Unsurprisingly, G-protein coupled receptors (GPCRs) and protein kinases dominate the data set, making up ~25% each of the total numbers of Clinical and Reference compounds. GPCRs are divided into three sub-classes, aminergic, peptidic and other, based on their endogenous agonists. It is clear from Table 2 that the kinases show only relatively minor NP enrichment in Clinical compounds, and the aminergic GPCRs virtually none. Clinical compounds acting at peptidic GPCRs in contrast have qualitatively higher % PNP and NP-likeness than their Reference compounds. Overall, there is no obvious link between the Clinical NP preferences of target classes and the numbers of Clinical compounds therein.

Post-2008 NP properties versus physicochemical properties

The NP results by target class were benchmarked by comparison with corresponding physical property trends. The results recapitulate the main conclusions of our earlier study using fewer Clinical compounds.[24] The most statistically significant and consistent differences between Clinical and Reference compounds by target class were carboaromatic ring count (lower in Clinical compounds), the fraction of carbon sp3 atoms (Fsp3) and the number of stereocenters (both increased in Clinical compounds). Shown in Figure 5 are the Fsp3 and carboaromatic ring results, together with the normalised spatial score (nSPS),[35] a measure of complexity that takes stereocenters and other topological features into account. Comparing Figures 4 and 5 indicate that the NP measures by target class are qualitatively similar to these physical properties as indicators of differences between Clinical versus Reference compounds.

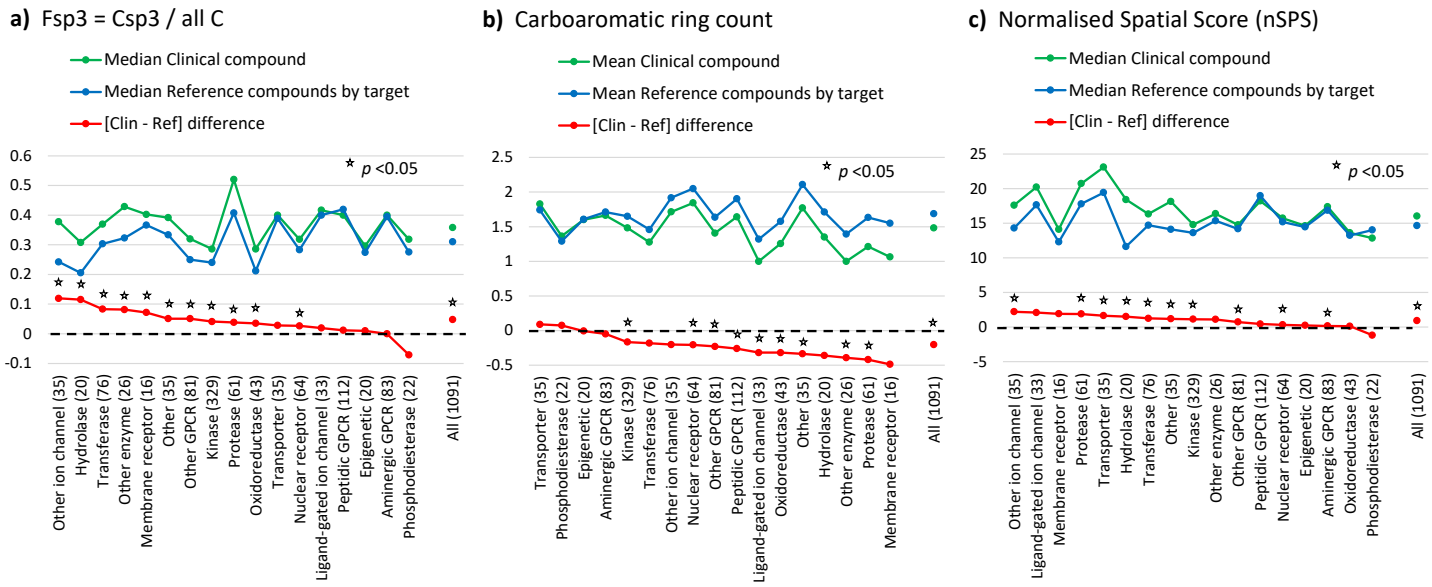

Figure 5. Benchmarked physical properties using the paired Clinical compound and Reference compounds by target set for comparison with the NP analysis in Figure 4. *p* Values are from Wilcoxon signed rank tests.

Fsp3 and stereocenter count are known to be greater in NP and naturally-inspired drugs, and aromatic ring count lower, versus synthetic drugs.[36,37] However, the [Clinical-Reference] differences for the NP properties of all clinical compounds (n=1091) are not significantly correlated with the corresponding values for these three properties, or for a range of other physicochemical properties (Table S.1). Similarly weak correlations are also seen using the target class data (n=17, not shown). This is further evidence that the three NP metrics can considered as independent measures of Clinical compound quality.

Table 3. Properties of post-2008 Clinical PNPs and NonPNPs.

Property*	PNP (n=666)	NonPNP (n=368)	Mean PNP versus Mean NonPNP	
	Mean, Median (Std. Dev.)		Difference	% Relative difference
NP fragments	3.17, 3 (1.01)	1.22, 1, (0.80)	2.05	159%
Frag_coverage_Murcko	0.74, 0.75 (0.18)	0.38, 0.38 (0.24)	0.36	93%
NP-like	-0.89, -0.97 (0.65)	-0.70, -0.85 (0.72)	-0.19	-27%
Carboaromatic rings	1.35, 1 (0.89)	1.71, 2 (0.86)	-0.36	-21%
Heteroaromatic rings	1.86, 2 (1.01)	0.73, 1 (0.72)	1.13	155%
Aromatic N atoms	2.80, 3 (1.72)	1.13, 1 (1.32)	1.67	149%
Carboaliphatic rings	0.47, 0 (0.85)	0.24, 0 (0.77)	0.23	97%
Heteroaliphatic rings	1.06, 1 (1.02)	0.66, 1 (0.77)	0.40	60%
Rotatable bonds	5.80, 5 (2.74)	6.93, 6 (3.79)	-1.13	-16%
H-bond acceptors	6.79, 7 (2.17)	5.36, 5 (2.19)	1.46	27%
PSA	98.7, 95.3 (34.6)	87.8, 88.4 (35.6)	10.9	12%
Mol. Wt, ALogP, cx_LogD, H-bond donors, stereocenters, Fsp3, pChEMBL, LE, LLE, QED				-10% to 10%

*Values shown are for properties that differ between PNPs and NonPNPs by >10%. ** % Relative difference = [(PNP/NonPNP)-1]x100 and for NP-like, [-(PNP/NonPNP)+1]x100. $p <$ 0.0001 in each case, by t-tests assuming unequal variance.

Consistent with their defined make-up, PNPs contain on average two more NP fragments than NonPNPs (illustrated for post-2008 Clinical compounds in Table 3; see Table S.2 for a full summary of the physical properties examined). Accordingly, PNPs and NonPNPs have markedly different ring counts, both aromatic and aliphatic. The number of heteroaromatic rings and aromatic nitrogen atoms are more than doubled in PNPs versus NonPNPs, while there are fewer carboaromatic rings in PNPs (Table 3). Aliphatic ring counts are also significantly increased in PNPs versus NonPNPs. Additionally, PNPs have one additional H-bond acceptor and one fewer rotatable bond versus NonPNPs. Notably, Fsp3 and numbers of stereocenters, which differ between Clinical and Reference compounds, are not different between PNPs and NonPNPs (Table S.2). This observation reinforces the independence of PNP fraction from these physical properties in distinguishing Clinical from Reference compounds.

All Clinical compounds (n=3173), contain 937 ring systems (using ring systems from DataWarrior's scaffold calculator[38]). New ring systems added per decade account for approximately 30% of all ring systems present, in agreement with a recent analysis.[39] However Clinical PNPs add more new ring systems than do Clinical NonPNPs, and the proportion of new ring systems is increasing over time in PNPs, but not NonPNPs (Figure S2). PNP ring system novelty versus NonPNPs is consistent with their higher aliphatic and heteroaromatic ring counts (Table 3).

NP fragments and fragment pair combinations present in Clinical and Reference compounds

Counts of all NP fragments, fragment pair combinations, and PNP fragment pair combinations occurring in the post-2008 Clinical and Reference compounds are summarised in Table 4. The fragment pair combinations can have more than one of each fragment type present. Of the 1673 individual fragments in the NP library, it is notable that only 176 (10.6%) are used in Clinical compounds, and a further 296 (17.7%) are unique to Reference compounds (Table 4). However, the 176 Clinical fragments comprise 97.8% of the total fragment coverage seen in Reference compounds. For the much higher numbers of fragment pair combinations, there is lower coverage by Clinical compounds (e.g. 72.8% for all fragment pairs and 63.1% for all fragment pair PNP combinations, Table 4) and Reference-only fragment pairs outnumber Clinical fragment pairs by

upto 10-fold across major target classes (e.g. GPCRs, Table 4). Reference singleton counts are dominant in both classes of fragment pairs.

Table 4. Total fragment counts in Clinical and Reference compounds for the post-2008 set.

Group	Major Target class*	n Clinical (% occurrence in Reference)	n Clinical only	n Clinical singletons	n Reference only (% occurrence)	n Reference singletons
Fragments	All	176 (97.8%)	2	53	296 (2.1%)	61
	GPCR	86 (93.5%)	1	26	197 (6.5%)	26
	Kinase	85 (96.8%)	1	26	178 (3.2%)	42
	Other	146 (95.4%)	1	46	277 (4.6%)	55
Fragment pair combinations	All	1074 (72.8%)	53	571	6372 (27.2%)	1855
	GPCR	318 (49.2%)	12	214	3542 (50.8%)	918
	Kinase	487 (66.7%)	14	276	2796 (33.3%)	839
	Other	669 (59.9%)	50	410	5093 (40.1%)	1566
Fragment pair PNP combinations	All	842 (63.1%)	67	534	5748 (36.9%)	1745
	GPCR	223 (39.0%)	21	169	2583 (61.0%)	802
	Kinase	363 (58.2%)	24	243	2447 (41.8%)	774
	Other	462 (45.9%)	47	324	4111 (54.1%)	1365

* Clinical and Reference compound counts: All, 1163 and 218773; GPCR, 284 and 59632; Kinase, 330 and 50370; Other, 574 and 115851.

The occurrence of the most common fragments and fragment combinations seen in all post-2008 Clinical compounds was compared to that seen in the Reference compounds. The 58 most common fragments, occurring ≥8 times in Clinical compounds (Figure 6), account for 90.5% of the total Clinical fragment occurrence. Of the 58 fragments, 15 (26%) occur significantly more frequently in Clinical versus Reference compounds (by odds ratio, $p < 0.05$), and 4 (7%) are significantly increased in Reference compounds (Figure 6). The 30 most common Clinical fragment pair combinations (Figure 7a), occurring ≥14 times, account for 22% of the total Clinical fragment occurrence and 10 of these (33%) are significantly more abundant clinically. Similar Clinical preferences are seen for the 31 most common PNP combinations, occurring ≥8 times and accounting for 22% of the total Clinical occurrence (Figure 7b). It is apparent from Figures 6 and 7 that the pyrrolidine and cyclopropyl rings feature prominently, individually and in combination with other fragments, as having higher abundance in Clinical over Reference compounds.

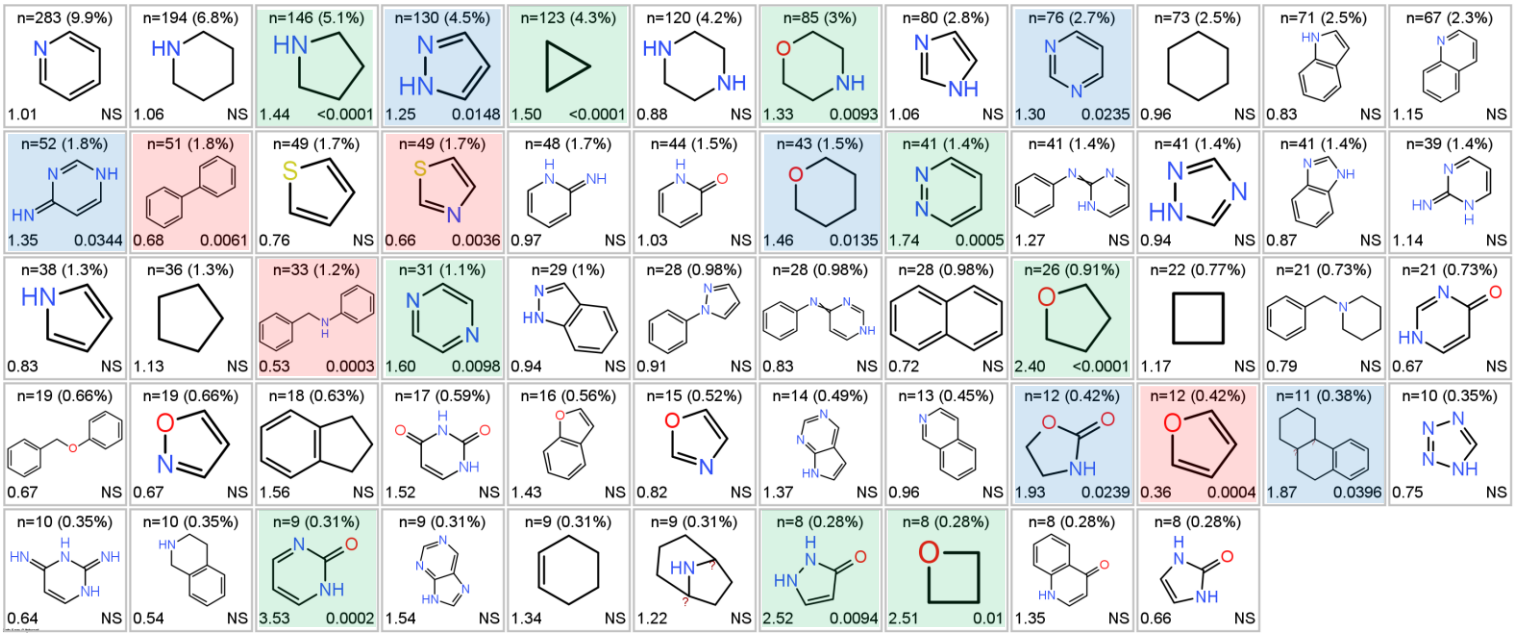

Figure 6. The 58 most abundant post-2008 Clinical NP fragments (≥ 8 occurrences, 90.5% of total Clinical NP fragment count). Shown for each fragment are: count (% clinical fragments) (top); odds ratio vs reference compounds (lower left); p value (lower right; NS = not significant, p >0.05). Clinical abundance increased (15, 26%): green, p <0.01; blue, p = 0.01-0.05). Clinical abundance decreased (4, 7%): red (pink, p <0.05). Canonical tautomers shown, as generated by RDKit.

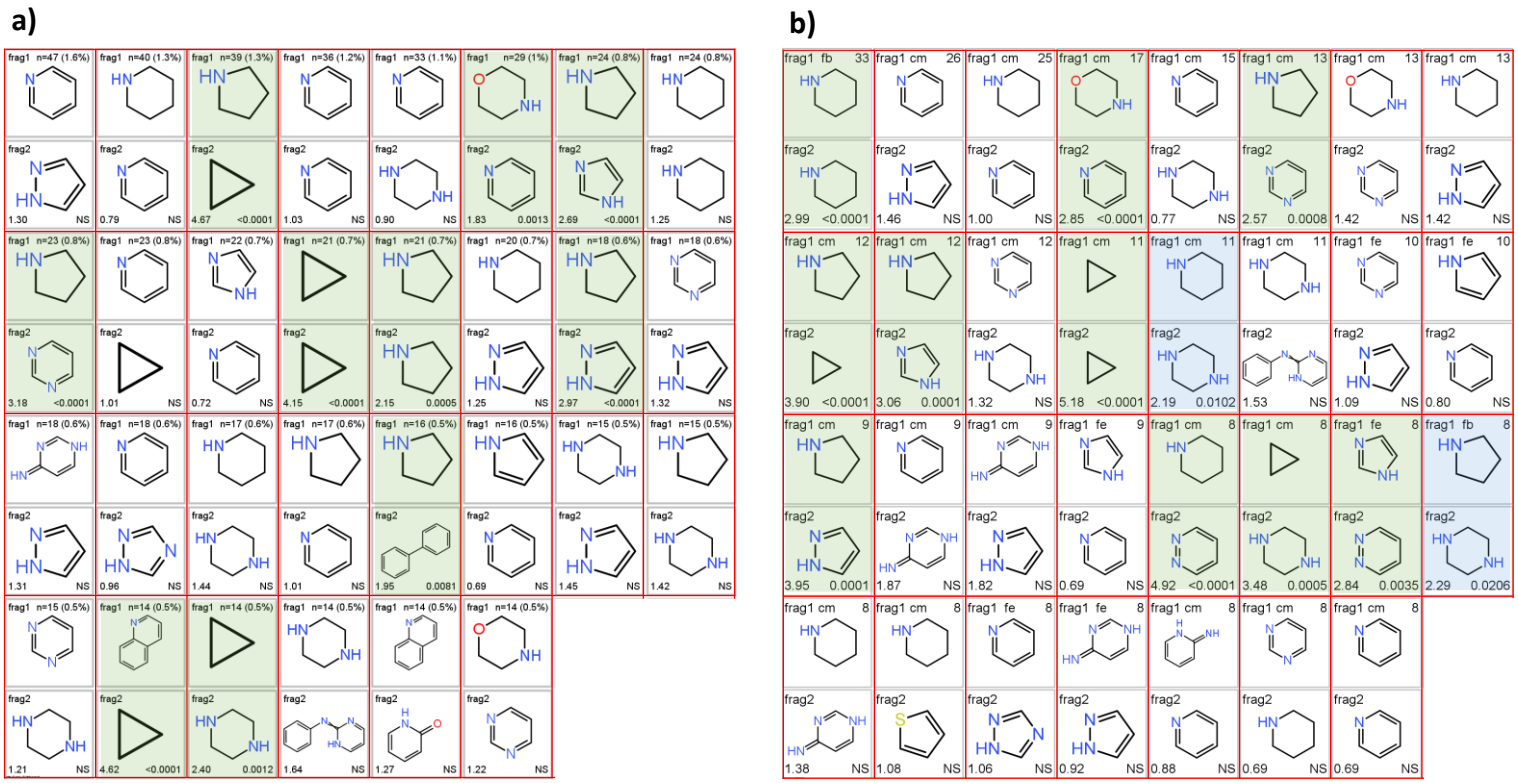

Figure 7. a) The 30 most abundant fragment pair combinations all Clinical compounds (≥14 Occurrences, 21.6% of all fragment pair occurrences). b) 31 Most abundant fragment combinations in all PNP Clinical compounds (≥8 occurrences, 22.0% of all PNP fragment pair occurrences). Shown are fragment pairs, count (%), odds ratio for Clinical vs Reference compounds and p value. Connectivity definitions (see ref. 16) between fragments in b) are: cm, connection monopodal (single bond connections); fb, bridge fused connections, 3-5 fused atoms; fe, fused edge connections, 2 fused atoms. Clinical abundance increased: green (p <0.01), blue (p< 0.01-0.05). Canonical tautomers shown, as generated by RDKit.

PNP trends over time in selected targets

The overall increase in PNP Clinical compounds (Figure 2a, b) is evident in heavily pursued targets that have seen long-term production of Clinical compounds. For the epidermal growth factor receptor-1 (EGFR-1) kinase (Figure 8), the invention of the 45 Clinical compounds found in ChEMBL has spanned 30 years. The first drug to the market, gefinitib, is not a PNP but contains 2 NP fragments, one of which, 4-anilinopyrimidine, provided inspiration for further development of the class, leading to approved PNPs such as lapatinib, followed later by brigatinib and osimertinib, which both contain a 2-anilinopyrimidine NP fragment (Figure 8). PNPs have been the dominant class of EGFR Clinical compounds since 2005, with their occurrence exceeding that seen in Reference compounds (Figure 8). Vascular endothelial growth factor receptor 2 (VEGF-2) kinase, with 54 Clinical candidates over 26 years, shows a similar increase In PNP Clinical compounds over time (Figure S3).

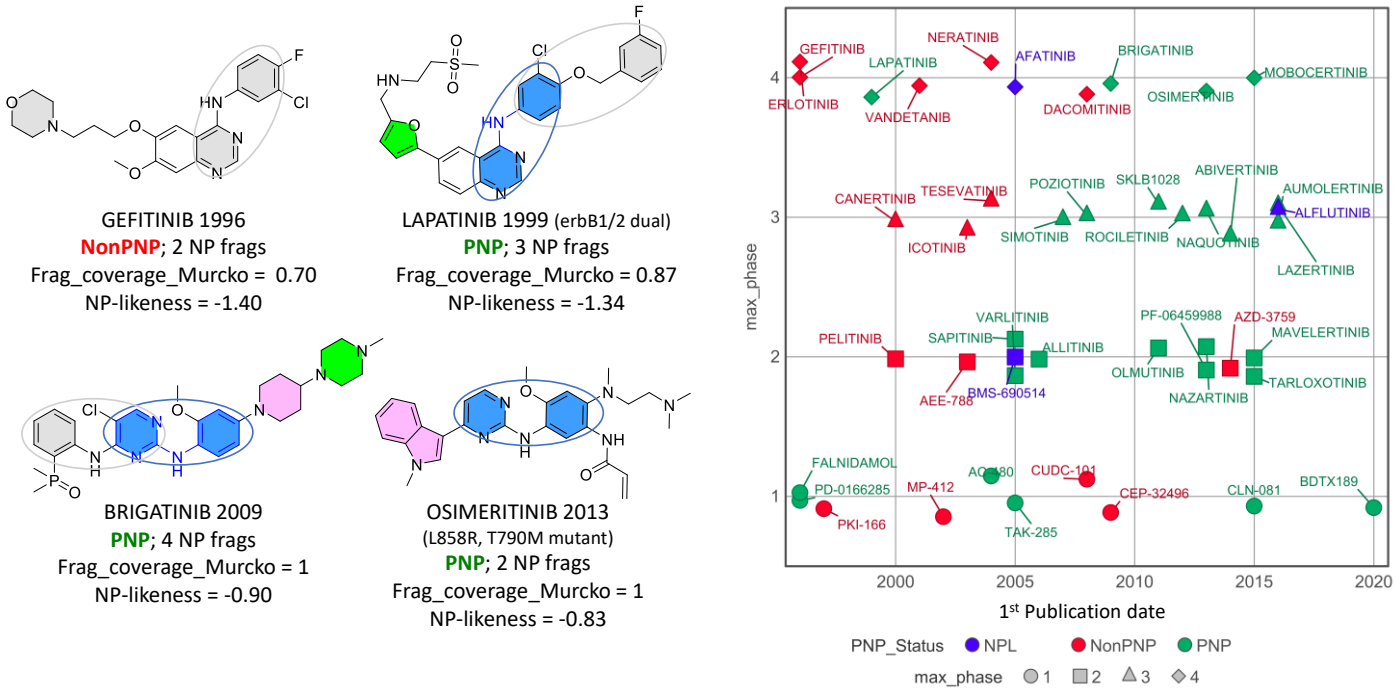

Figure 8. Development of EGFR1 (erbB1) Antagonists. Colour filled NP fragments show molecules classified as pseudo-natural products (PNP). In all, 26 of 44 Clinical compounds since 1995 are PNPs (59%) and 2899 of 5480 Reference compounds are PNPs (53%); odds ratio = 1.29 (p >0.05)). Since 2005, 22 of 29 Clinical compounds are PNPs (76%) and 1499 3193 Reference compounds are PNPs (47%); odds ratio = 3.65 (p=0.0036).

Unsurprisingly, amongst targets having multiple Clinical compounds, exploitation of the first molecules discovered is common practice, leading to subsequent molecules that often share

comparable pharmacophoric features, a tendency that can also lead to promulgation of the same PNP_Status. An example is the neurokinin-1 (NK-1) receptor, a peptidic GPCR, which provided 16 Clinical compounds in the ChEMBL database (with five marketed, including two prodrugs) disclosed between 1992 and 2007. All NK-1 antagonists that reached Phase 3, and all discovered after 2002, are classified NPL or PNP (Figure 9). Following aprepitant in 1995, all subsequent NK-1 antagonists employ a common 3,5-difluoromethylphenyl group and an additional phenyl ring. Aprepitant and netupitant bind to NK-1 in a similar mode,[40] except for the pendant heterocycles that are part of their PNP motifs (triazolone and piperazine respectively, Figure 9). However, both compounds are biologically comparable since they induce an identical intra-helical H-bonding network in the NK-1 structure, which might explain the advantageous insurmountable antagonism they exhibit.[40] The NK-1 receptor stands out as a target where PNP occurrence in Clinical compounds (10/16, 63%) significantly exceeds than seen in Reference compounds (470/2158, 22%), with an odds ratio of 5.99 (p = 0.0006).

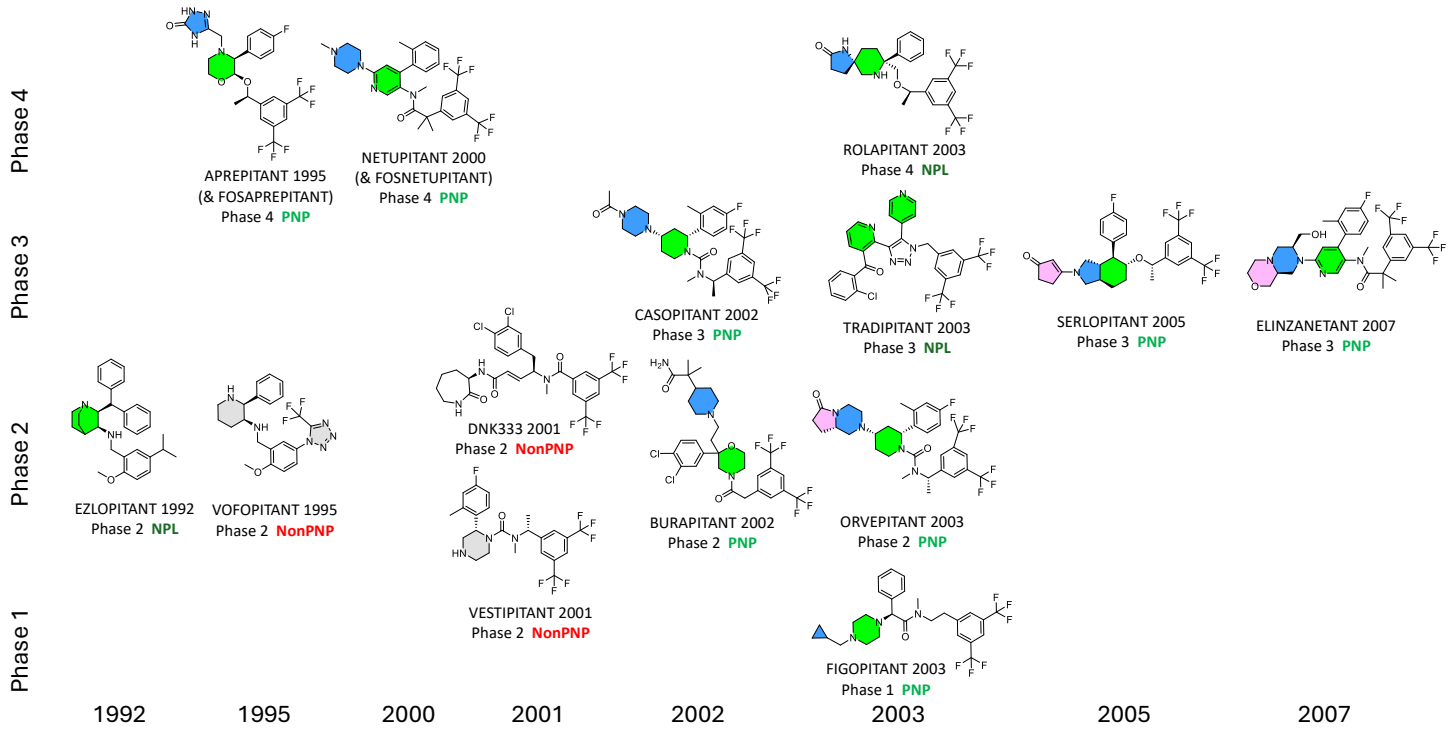

Figure 9. Development of Clinical NK-1 antagonists. Colour filled NP fragments show molecules classified as pseudo-natural products (PNP) or NP-like (NPL); PNP occurrence in Clinical compounds (10/16, 63%) significantly exceeds that seen in Reference compounds (470/2158, 22%), with an odds ratio of 5.99 (p = 0.0006).

The peroxisome proliferator-activated receptor (PPAR) subtypes, where PPARα and/or PPARγ agonists have been pursued for diabetes, are examples of targets that have been less responsive to

application of NP fragments. From 1990 to 2007, PNP occurrence in Clinical compounds active at the PPARγ subtype, (3/23, 14%) is similar to that seen in the corresponding Reference compounds (786/3564, 22%). Ten PPARγ Clinical compounds since 1990 have reached Phase 3, with two, saroglitizar and lobeglitazone, marketed in India and South Korea respectively (Figure 10). These molecules are clearly follow-ons from marketed thiazolidinones discovered before 1990, such as rosiglitazone and troglitazone (Figure 10), which suffered from cardiovascular safety issues typically encountered in this class.

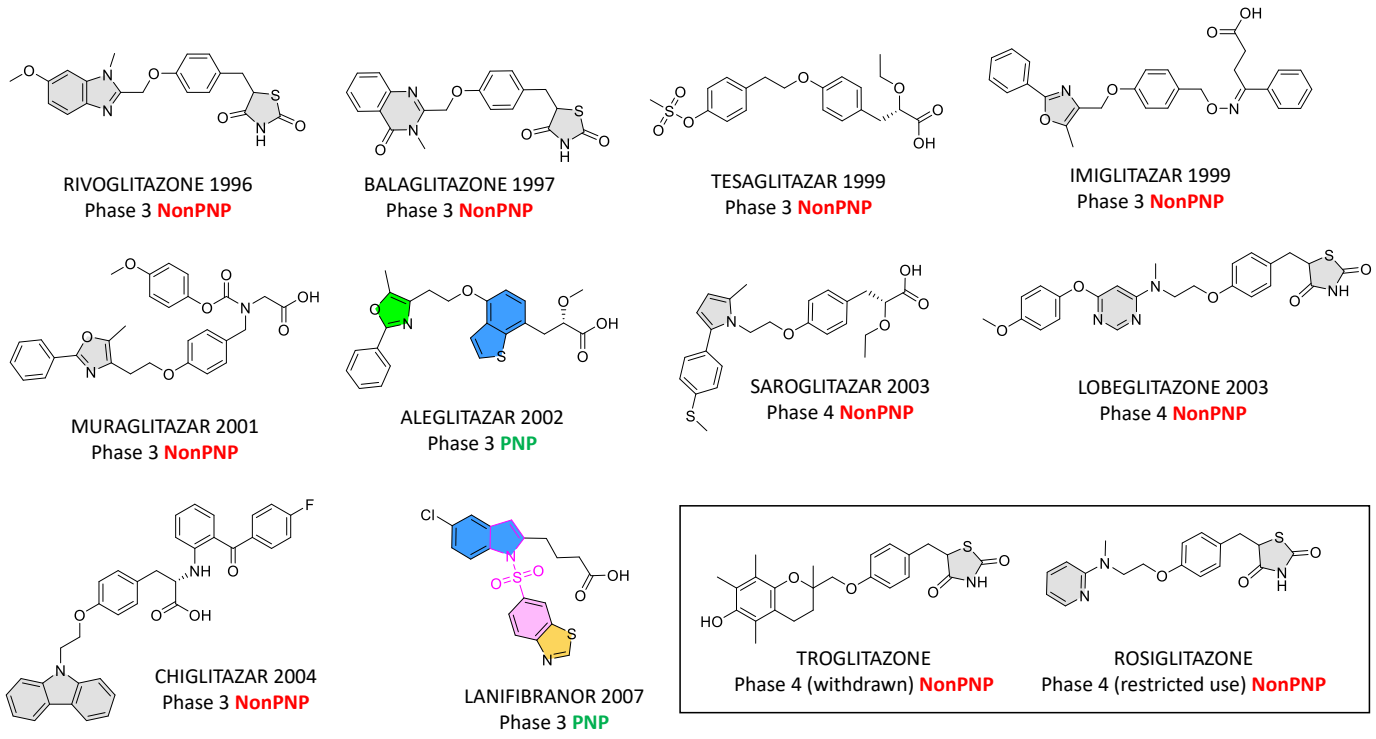

Figure 10. Development of Phase 3 and Phase 4 Clinical PPAR agonists, where NP fragments have been infrequently used. Colour filled NP fragments show molecules classified as pseudo-natural products (PNP). Saroglitazar (approved in India) is reported as Phase 3 in ChEMBL and lobeglitazone (approved in South Korea) was not present in ChEMBL Clinical compounds. PNP occurrence in all Clinical compounds active at the PPARγ subtype is 3/23 (14%) and for Reference compounds 786/3564 (22%).

Summary and Discussion

We have used three measures of natural product (NP) character to assess compounds reaching the Clinic (phases 1-4), as well as Reference compounds acting at the same biological targets. While calculated NP-likeness[27] is in overall decline in drug discovery and in Clinical compounds, the results show that the 'NP signal' remains strong because of widespread exploitation of a relatively small pool of NP fragments. Notably, pseudo-natural products (PNPs), which contain NP fragments but are inaccessible by Nature's known biosynthetic pathways, have been increasing progressively over time and make up 67% of Clinical compounds first disclosed since 2010. Inevitably the growth of PNPs in Clinical molecules cannot continue at the rate observed and a levelling off will occur; the 2021 year data (Figure 3a) may indicate this is underway. The overall decline in NP-likeness of Clinical compounds can be attributable mainly to an increase in aromatic ring count together with a decrease in oxygen atom count.[4, 36] PNPs have introduced more ring system novelty to the clinic versus NonPNPs (Figure S2) while having lower[16] NP-likeness (Table 3).

In support of a 'natural selection' process occurring in successful drug discovery, the results (Figures 3, 4, Table 2) show that since 2008, Clinical compounds have increased NP property values versus target-matched Reference compounds. Overall in this period, PNPs are 54% more likely to be found in Clinical versus Reference compounds. On the other hand, not all target classes show a NP signal, there is less evidence for a NP property clinical increase prior to 2008, and Clinical and Reference compounds have the same relative abundance in the NPL (NP-like) class in all time frames (Figure 3). What has caused the overall separation of PNP abundance in Clinical and Reference compounds published after 2008? Developments in the late 1990s to early 2000s such as the maturity of compound screening collection enhancements and multiparameter lead optimisation, may be contributory factors. In the same time frame, the increased attention paid to physical property control is an additional factor: higher Fsp3 and complexity, with lower carboaromatic ring count are evident in Clinical compounds versus Reference compounds (Figure 5) and these trends are consistent with increased NP character.

Fourteen of the seventeen target classes show NP Clinical enhancement to varying degrees, including proteases, nuclear receptors and oxidoreductases (Table 2). In contrast, aminergic G-protein coupled receptors (GPCRs), display no Clinical compound NP enhancement, while other GPCRs and protein kinases have only marginal Clinical NP enhancement across the categories in Table 2. Can the diminished Clinical NP enhancements in these major classes (45% of post-2008 Clinical compounds) be explained? One possible factor at play might be the historically long-established and highly exemplified options for the essential pharmacophores seen in both aminergic

GPCR and kinase competitive inhibitors. The majority of kinase inhibitors are competitive with ATP, and kinase drug fragments have been reported[41] to cover a very small fraction of possible hinge-binder[42] space. Kinase inhibitors frequently bind to multiple other kinases[43] and existing kinase inhibitor libraries are commonly used to identify new kinase leads by focussed screening.[44,45] Competitive aminergic GPCR antagonists have historically typically contained a common aromatic ring and a basic group,[46] as found in endogenous agonists (e.g. dopamine, serotonin, histamine, noradrenaline), and can employ similar ligand-protein interactions within each receptor sub-class.[47] Targets amongst the protease, oxidoreductase and nuclear receptor (including nuclear hormone receptors and transcription factors) families, which all show Clinical compound NP enhancement, are more diverse in their substrate requirements.

Much current drug discovery uses known structures as starting points,[2,3] and exploiting biologically proven or 'privileged' structures is a prominent strategy used in developing screening collections. 'Recycling' of knowledge is clearly a major factor influencing the NP profiles of Clinical candidates, as shown by the target examples (Figures 8-10). Most striking is the fact that as few as 176 NP fragments account for, on average, 63% of the heavy (non-H) atoms in the Murcko scaffold structures of Clinical compounds disclosed since 2008. Further, just 58 NP fragments (Figure 6) cover 90.5% of all Clinical compound NP fragment space. We therefore extend the questions already raised regarding the necessity for novel ring systems in drugs[39,48] to ask if application of NP fragments is a preferred strategy for success, instead of hunting for diversity amongst the vast numbers of less-used and virtual non-NP ring systems and scaffolds.[49,50]

The rapid growth in Clinical PNPs has probably occurred largely unknowingly because the PNP concept, disclosed in 2020,[12] is too new to have had significant impact on Clinical compound design. The observed PNP and NP fragment density increases may have mostly occurred intuitively, rather than by 'NP design', because the most frequently used NP fragments (Figure 6) are all very well-known to, and commonly employed by, medicinal chemists. One example of the application of an NP moiety in lead optimisation is the use of *R*-(+)-carvone in the discovery of the JAK kinase inhibitor tofacitinib,[51] itself a PNP.

The observed Clinical NP fragment preferences in the data set (Figures 6 and 7) are intriguing and suggest specific fragments and combinations that could take priority in compound design. An important point is that any specific NP fragments contained in Clinical compounds which were generally introduced at late stages in the lead optimisation processes, could appear 'enhanced' in Clinical versus comparative Reference compounds. An example is oxetane, a relatively new entry to the repertoire[52] which is enriched in Clinical compounds (Figure 6, 8th entry in 5th row). This effect

would be less likely to be seen for NP fragments present in starting hit structures. Quantifying the impact of these aspects is not practicable in the current dataset. The narrowing of successful design options using specific NP fragments, which occurs as lead optimisation progresses, is consistent with a 'natural selection' hypothesis.

There is a significant pool of under-used NP fragments, and an immense number of novel NP fragment combinations are available to exploit to create NP-biased screening libraries and building blocks for use in lead optimisation. The design principles for PNP molecules have been laid out[12-14, 16, 18, 19] and are ready for exploitation by medicinal chemists. PNPs provide greater ring system novelty in the clinic and predictably distinctive physicochemical properties, especially increased nitrogenous aromatic rings, versus NonPNPs. However, it is our experience that recognising a PNP structure and knowing those NP fragment combinations that are non-biosynthetic, are often not intuitive or straightforward. Applying the available PNP algorithm[17] in advance of synthesis is a necessary step. In designing new PNPs, synthetic challenges are likely to arise – the investment required needs to be balanced against the accumulating evidence supporting their biological relevance, clinical dominance and untapped potential. However, we note that nowadays thousands of PNPs have become available from commercial sources, such that PNP libraries can readily be established without extensive synthesis efforts.[17] Further extension to consider additional NP ring systems and acyclic frameworks[29,30] is also possible.

A limitation to this study surrounds the inevitably incomplete nature of freely available Reference compound datasets. To help corroborate our results, we encourage owners of complete drug discovery project databases to examine the NP properties of optimised leads and candidate drugs versus others synthesised, an approach that does not require disclosure of proprietary structures in publication.[53] Similarly, higher NP character of identified hits versus compound screening collections, if seen, could help shed light on the presence of 'dark' chemical matter.[54] Nevertheless, we believe that the results showing Clinical NP enrichment based on published compound data are sufficiently positive to warrant a greater focus on the application of NP properties and fragments in compound design.

Conclusions

By analysing data published in journals, extracted from ChEMBL version 32,[22] we provide evidence that the natural product (NP) properties of a set of >1000 Clinical compounds published since 2008 are increased versus corresponding time and target-matched Reference compounds. The magnitude of the NP enhancement seen varies by target class, with positive a signal in seen in the majority.

Kinases and aminergic GPCRs, highly explored target classes with a long history and limited endogenous ligand diversity, show weak or no NP enhancement. The increase in Clinical compound NP properties results from the use of just 176 NP fragments, which together make up, on average, 63% of the heavy atoms in post-2008 Clinical compound core scaffolds. Clinical pseudo-natural products (PNPs), where two or more NP fragments are combined in ways not achievable in Nature, have been rapidly increasing over time and comprise 67% of post-2010 Clinical compounds. The overall results are supportive of the occurrence of 'natural selection' being associated with many successful drug discovery campaigns. It has been proposed that NP-likeness assists drug distribution by membrane transporters[21] and we further speculate that employing NP fragments may result in less attrition due to toxicity, a major cause of preclinical failure.[55]

There is huge untapped potential in further exploitation of currently used and unused NP fragments, especially in fragment combinations and design of PNPs, without the need to resort to chemically diverse ring systems and scaffolds. To exploit these opportunities, it is vital that 'NP awareness' is added to the repertoire of medicinal chemists. Sir James Black, discoverer of propranolol and cimetidine, famously stated that "the most fruitful basis for the discovery of a new drug is to start with an old drug."[56] In the genomic era, the Black principle holds ever true, as medicinal chemists, knowingly or unknowingly, are repeatedly using a small group of established NP fragments to discover clinical candidates. We concur with the sentiment that "the local chemical space of a natural product can prove superior to the natural product itself."[57] Adding NP fragment-based parameterisation to enhance machine learning models and influence generative chemistry is recommended.

NP structural motifs are provided pre-designed by Nature, constructed for biological purposes as a result of 4 billion years of evolution. In short, applying Nature's building blocks - Natural Intelligence - to drug design can enhance the opportunities now offered by Artificial Intelligence.

Supporting information

Figures S1 – S3. NP metric Clinical versus Reference compounds by target class using unpaired data for all compounds and targets with ≥100 reference compounds; ring system novelty in PNPs and NonPNPs over time in clinical compounds; development of VEGF inhibiors.
Tables S1 – S2. Cross correlation of NP measures versus other properties for post-2008 Clinical compounds; physical properties of post-2008 Clinical compounds by PNP_Status.
Excel spreadsheet S1. NP fragment list; post-2008 Clinical compound-target pairs; post-2008 Reference compounds by target; statistical data for Figures.

Acknowledgements

We acknowledge funding from the Member States of the European Molecular Biology Laboratory and the Wellcome Trust [104104/A/14/Z, 218244/Z/19/Z, 228142/Z/23/Z]. Funding for open access charge: Wellcome Trust. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission. P. D. L. thanks the University of Nottingham for the provision of Scifinder®.

References

1. Koehn, F. E.; Carter, G. T. The evolving role of natural products in drug discovery. *Nat. Rev. Drug Discov.* **2005***, 4*, 206–220.
2. Brown, D. G.; Boström, J. Where Do Recent Small Molecule Clinical Development Candidates Come From? *J. Med. Chem*. **2018**, *61*, 9442−9468.
3. Brown, D. G. An Analysis of Successful Hit-to-Clinical Candidate Pairs. *J. Med. Chem*. **2023**, *66*, 7101−7139.
4. Young, R. J.; Flitsch, S. L.; Grigalunas, M.; Leeson, P. D.; Quinn, R. J.; Turner, N. J.; Waldmann, H. The Time and Place for Nature in Drug Discovery. *JACS Au.* **2022**, *2*, 2400–2416.
5. Newman, D. J.; Cragg, G. M. Natural Products as Sources of New Drugs over the Nearly Four Decades from 01/1981 to 09/2019. *J. Nat. Prod.* **2020**, *83*, 770–803.
6. Rodrigues, T.; Reker, D.; Schneider, P.; Schneider, G. Counting on natural products for drug design. *Nat. Chem.* **2016**, *8*, 531–541.
7. Harvey, A. L.; Edrada-Ebel, R.; Quinn, R. J. The re-emergence of natural products for drug discovery in the genomics era. *Nat. Rev. Drug Discov.* **2015**, *14*, 111–129.
8. Huffman, B. J.; Shenvi, R. A. Natural Products in the “Marketplace”: Interfacing Synthesis and Biology. *J. Am. Chem. Soc.* **2019**, *141*, 3332-3346.
9. Pye, C. R.; Bertin, M. J.; Lokey, R. S.; Gerwick, W. H.; Linington, R. G. Retrospective analysis of natural products provides insights for future discovery trends. *Proc. Natl Acad. Sci. USA* **2017**, *114*, 5601-5606.
10. Atanasov, A. G.; Zotchev, S. B.; Dirsch, V. M.; Orhan, I. E.; Banach, M.; Rollinger, J. M.; Barreca, D.; Weckwerth, W.; Bauer, R.; Bayer, E. A.; Majeed, M.; Bishayee, A.; Bochkov, V.; Bonn, G. K.; Braidy, N.; Bucar, F.; Cifuentes, A.; D’Onofrio, G.; Bodkin, M.; Diederich, M.; Dinkova-Kostova, A. T.; Efferth, T.; El Bairi, K.; Arkells, N.; Fan, T. P.; Fiebich, B. L.; Freissmuth, M.; Georgiev, M. I.; Gibbons, S.; Godfrey, K. M.; Gruber, C. W.; Heer, J.; Huber, L. A.; Ibanez, E.; Kijjoa, A.; Kiss, A. K.; Lu, A.; Macias, F. A.; Miller, M. J. S.; Mocan, A.; Müller, R.; Nicoletti, F.; Perry, G.; Pittalà,

V.; Rastrelli, L.; Ristow, M.; Russo, G. L.; Silva, A. S.; Schuster, D.; Sheridan, H.; Skalicka-Woźniak, K.; Skaltsounis, L.; Sobarzo-Sánchez, E.; Bredt, D. S.; Stuppner, H.; Sureda, A.; Tzvetkov, N. T.; Vacca, R. A.; Aggarwal, B. B.; Battino, M.; Giampieri, F.; Wink, M.; Wolfender, J. L.; Xiao, J.; Yeung, A. W. K.; Lizard, G.; Popp, M. A.; Heinrich, M.; Berindan-Neagoe, I.; Stadler, M.; Daglia, M.; Verpoorte, R.; Supuran, C. T. Natural Products in Drug Discovery: Advances and Opportunities. *Nat. Rev. Drug Discov.* **2021**, *20*, 200–216.

11. a) Mullowney, M. W.; Duncan, K. R.; Elsayed, S. S.; Garg, N.; van der Hooft, J. J. J.; Martin, N. I.; Meijer, D.; Terlouw, B. R.; Biermann, F.; Blin, K.; Durairaj, J.; Gorostiola González, M.; Helfrich, E. J. N.; Huber, F.; Leopold-Messer, S.; Rajan, K.; de Rond, T.; van Santen, J. A.; Sorokina, M.; Balunas, M. J.; Beniddir, M. A.; van Bergeijk, D. A.; Carroll, L. M.; Clark, C. M.; Clevert, D.-A.; Dejong, C. A.; Du, C.; Ferrinho, S.; Grisoni, F.; Hofstetter, A.; Jespers, W.; Kalinina, O. V.; Kautsar, S. A.; Kim, H.; Leao, T. F.; Masschelein, J.; Rees, E. R.; Reher, R.; Reker, D.; Schwaller, P.; Segler, M.; Skinnider, M. A.; Walker, A. S.; Willighagen, E. L.; Zdrazil, B.; Ziemert, N.; Goss, R. J. M.; Guyomard, P.; Volkamer, A.; Gerwick, W. H.; Kim, H. U.; Müller, R.; van Wezel, G. P.; van Westen, G. J. P.; Hirsch, A. K. H.; Linington, R. G.; Robinson, S. L.; Medema, M. H. Artificial Intelligence for Natural Product Drug Discovery. *Nat. Rev. Drug Discov*. **2023**, *22*, 895-916. b) Saldivar-Gonzalez, F. I.; Aldas-Bulos, V. D.; Medina-Franco, J. L.; Plisson, F. Natural product drug discovery in the artificial intelligence era. *Chem. Sci.* **2022**, *13*, 1526–1546.
12. Karageorgis, G.; Foley, D. J.; Laraia, L.; Waldmann, H. Principle and design of pseudo-natural products. *Nat. Chem.* **2020**, *12*, 227– 235.
13. Karageorgis, G.; Foley, D. J.; Laraia, L.; Brakmann, S.; Waldmann, H. Pseudo Natural Products - Chemical Evolution of Natural Product Structure. *Angew. Chem., Int. Ed*. **2021**, *60*, 15705.
14. Grigalunas, M.; Brakmann, S.; Waldmann, H. Chemical Evolution of Natural Product Structure. *J. Am. Chem. Soc*. **2022**, *144*, 3314−3329.
15. Over, B.; Wetzel, S.; Grütter, C.; Nakai, Y.; Renner, S.; Rauh, D.; Waldmann, H. Natural-product-derived fragments for fragment-based ligand discovery. *Nat. Chem.* **2013,** *5*, 21–28.
16. Gally, J.-M.; Pahl, A.; Czodrowski, P.; Waldmann, H. Pseudonatural Products Occur Frequently in Biologically Relevant Compounds. *J. Chem. Inf. Model.* **2021**, *61*, 5458–5468.
17. Pahl, A.; Grygorenko, O. O.; Kondratov, I. S.; Waldmann, H. Identification of Readily Available Pseudo-Natural Products. *ChemRxiv* **2024**. Manuscript in preparation.
18. Grigalunas, M.; Burhop, A.; Zinken, S.; Pahl, A.; Gally, J.-M.; Wild, N.; Mantel, Y.; Sievers, S.; Foley, D. J.; Scheel, R.; Strohmann, C.; Antonchick, A. P.; Waldmann, H. Natural Product Fragment Combination to Performance-Diverse Pseudo-Natural Products. *Nat. Commun*. **2021**, *12*, 1883.
19. Liu, J.; Grigalunas, M.; Waldmann, H Chemical evolution of natural product structure for drug discovery. *Ann. Rep. Med. Chem.* **2023**, *61*, 1-53.

20. a) O'Hagan, S.; Swainston, N.; Handl, J.; Kell, D. B. A 'rule of 0.5' for the metabolite-likeness of approved pharmaceutical drugs. *Metabolomics* **2015**, *11*, 323−339. b) Kell, D.B. Implications of endogenous roles of transporters for drug discovery: hitchhiking and metabolite-likeness. *Nat. Rev. Drug Disc.* **2016,** 15, 143. c) Kell, D. B. The Transporter-Mediated Cellular Uptake and Efflux of Pharmaceutical Drugs and Biotechnology Products: How and Why Phospholipid Bilayer Transport Is Negligible in Real Biomembranes. *Molecules* **2021**, *26*, 5629.
21. O'Hagan, S.; Kell, D. B. Consensus rank orderings of molecular fingerprints illustrate the most genuine similarities between marketed drugs and small endogenous human metabolites, but highlight exogenous natural products as the most important 'natural' drug transporter substrates. *ADMET and DMPK* **2017**, *5*, 85−125.
22. Heinzke, A. L.; Zdrazil, B.; Leeson, P. D.; Young, R. J.; Pahl, A.; Waldmann, H.; Leach,' A. R. A compound-target pairs dataset: differences between drugs, clinical candidates and other bioactive compounds. *ChemRxiv*, **2024**. Doi: 10.26434/chemrxiv-2024-vj70m-v2
23. Zdrazil, B.; Felix, E.; Hunter, F.; Manners, E. J.; Blackshaw, J.; Corbett, S.; de Veij, M.; Ioannidis, H.; Lopez, D. M.; Mosquera, J. F.; Magarinos, M. P.; Bosc, N.; Arcila, R.; Kiziloren, T.; Gaulton, A.; Bento, A. P.; Adasme, M. F.; Monecke, P.; Landrum, G. A.; Leach, A. R. The ChEMBL Database in 2023: A Drug Discovery Platform Spanning Multiple Bioactivity Data Types and Time Periods. *Nucleic Acids Res.* **2023**, *52*, D1180−D1192.
24. Leeson, P. D.; Bento, A. P.; Gaulton, A.; Hersey, A.; Manners, E. J.; Radoux, C. J.; Leach, A. R. Target-Based Evaluation of "Drug-Like" Properties and Ligand Efficiencies. *J. Med. Chem*. **2021**, *64*, 7210-7230.
25. a) Leeson, P. D.; Davis, A. M. Time-related differences in the physical property profiles of oral drugs. J*. Med. Chem*. **2004**, *47*, 6338-6348. b) Shultz, M. D. Two Decades under the Influence of the Rule of Five and the Changing Properties of Approved Oral Drugs. *J. Med. Chem*. **2019**, *62*, 1701−1714. c) Agarwal, P.; Huckle, J.; Newman, J.; Reid, D. L. Trends in Small Molecule Drug Properties: A Developability Molecule Assessment Perspective. *Drug Discov. Today* **2022**, *27*, 103366.
26. Bemis, G. W.; Murcko, M. A. The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem*. **1996**, *39*, 2887–2893.
27. Ertl, P.; Roggo, S.; Schuffenhauer, A. Natural Product-likeness Score and Its Application for Prioritization of Compound Libraries. *J. Chem. Inf. Model.* **2008**, *48*, 68–74.
28. a) The RDKit Documentation — The RDKit 2023.09.6 documentation. b) https://zenodo.org/records/10099869).
29. Ertl, P. Substituents of life: The most common substituent patterns present in natural products. *Bioorg. Med. Chem.* **2022**, *54*, 116562.

30. Chen, Y.; Rosenkranz, C.; Hirte, S.; Kirchmair, J. Ring systems in natural products: structural diversity, physicochemical properties, and coverage by synthetic compounds. *Nat. Prod. Rep.* **2022**, *39*, 1544-1556.
31. Langdon, S. R.; Brown, B.; Blagg, J. Scaffold Diversity of Exemplified Medicinal Chemistry Space. *J. Chem. Inf. Model.* **2011**, *51*, 2174–2185.
32. Zdrazil, B.; Guha, R. The rise and fall of a scaffold: a trend analysis of scaffolds in the medicinal chemistry literature. *J. Med. Chem.* **2018**, *61*, 4688–4703.
33. Landrum, G. A.; Riniker, S. Combining IC50 or *Ki* Values from Different Sources Is a Source of Significant Noise. *J. Chem. Inf. Model.* **2024**, *64*, 1560–1567.
34. Morphy, R. The Influence of Target Family and Functional Activity on the Physicochemical Properties of Pre-Clinical Compounds. *J. Med. Chem*. **2006**, *49*, 2969-2978.
35. Krzyzanowski, A.; Pahl, A.; Grigalunas, M.; Waldmann, H. Spacial Score - A Comprehensive Topological Indicator for Small-Molecule Complexity. *J. Med. Chem*. **2023**, *66*, 12739-12750.
36. Stratton, C. F.; Newman, D. J.; Tan, D. S. Cheminformatic comparison of approved drugs from natural product versus synthetic origins, *Bioorg. Med. Chem. Lett*. **2015**, *25*, 4802–4807.
37. Leeson, P. D. Molecular inflation, attrition and the rule of five. *Adv. Drug Del. Rev*. **2016**, *101*, 22-33.
38. a) Sander, T.; Freyss, J.; von Korff, M.; Rufener, C. DataWarrior: An Open-Source Program For Chemistry Aware Data Visualization And Analysis. *J. Chem. Inf. Model.* **2015**, *55*, 460–473. b) www.openmolecules.org
39. Shearer, J.; Castro, J. L.; Lawson, A. D.; MacCoss, M.; Taylor, R. D. Rings in clinical trials and drugs: Present and future. *J. Med. Chem*. **2022**, *65*, 8699-8712.
40. Schöppe, J.; Ehrenmann, J.; Klenk, C.; Rucktooa, P.; Schütz, M.; Doré. A. S.; Plückthun, A. Crystal structures of the human neurokinin 1 receptor in complex with clinically used antagonists. *Nat. Commun*. **2019,** *10*, 17.
41. Zhao, H.; Caflisch, A. Current kinase inhibitors cover a tiny fraction of fragment space. *Bioorg. Med. Chem. Lett*. **2015**, *25*, 2372–2376.
42. Zhao, Z.; Bourne, P. E. How Ligands Interact with the Kinase Hinge. *ACS Med. Chem. Lett.* **2023**, *14*, 1503–1508.
43. Klaeger, S.; Heinzlmeir, S.; Wilhelm, M.; Polzer, H.; Vick, B.; Koenig, P.-A.; Reinecke, M.; Ruprecht, B.; Petzoldt, S.; Meng, C.; Zecha, J.; Reiter, K.; Qiao, H.; Helm, D.; Koch, H.; Schoof, M.; Canevari, G.; Casale, E.; Depaolini, S. R.; Feuchtinger, A.; Wu, Z.; Schmidt, T.; Rueckert, L.; Becker, W.; Huenges, J.; Garz, A.-K.; Gohlke, B.-O.; Zolg, D. P.; Kayser, G.; Vooder, T.; Preissner, R.; Hahne, H.; Tõnisson, N.; Kramer, K.; Götze, K.; Bassermann, F.; Schlegl, J.; Ehrlich, H.-C.; Aiche, S.; Walch, A.; Greif, P. A.; Schneider, S.; Felder, E. R.; Ruland, J.; Médard, G.; Jeremias, I.; Spiekermann, K.; Kuster, B. The target landscape of clinical kinase drugs. *Science* **2017**, *358*, eaan4368.

44. Lightfoot. H. L.; Goldberg, F. W.; Sedelmeier, J. Evolution of Small Molecule Kinase Drugs. *ACS Med. Chem. Lett*. **2019**, *10*, 153−160.
45. Kettle, J. G.; Wilson, D. M. Standing on the shoulders of giants: a retrospective analysis of kinase drug discovery at AstraZeneca. *Drug Discov. Today* **2016**, *21*, 1596−1608.
46. Lloyd, E. J.; Andrews, P. R. A Common Structural Model for Central Nervous System Drugs and Their Receptors *J. Med. Chem*. **1986**, *29*, 453–462.
47. Vass, M.; Podlewska. S.; de Esch, I. J. P.; Bojarski, A. J.; Leurs, R.; Kooistra. A. J.; de Graaf, C. Aminergic GPCR–Ligand Interactions: A Chemical and Structural Map of Receptor Mutation Data. *J. Med. Chem.* **2019**, *62*, 3784–3839.
48. Taylor, R. D.; MacCoss, M.; Lawson, A. D, Combining Molecular Scaffolds from FDA Approved Drugs: Application to Drug Discovery. *J. Med. Chem.* **2017**, *60*, 1638–1647.
49. a) Ertl, P. Magic Rings: Navigation in the ring chemical space guided by the bioactive rings. *J. Chem. Inf. Model.* **2021**, *62*, 2164-2170. b) Ertl, P.; Altmann, E.; Racine, S.; Lewis, R. Ring replacement recommender: Ring modifications for improving biological activity. *Eur. J. Med. Chem.* **2022**, 114483. c) Ertl, P. Database of 4 Million Medicinal Chemistry-Relevant Ring Systems. *J. Chem. Inf. Model.* **2024**, *64*, 1245–1250.
50. a) Visini, R.; Arús-Pous, J. ; Awale, M. ; Reymond, J-L. Virtual Exploration of the Ring Systems Chemical Universe. *J. Chem. Inf. Model*. **2017**, *57*, 2707–2718. b) Ye Buehler. Y.; Reymond, J-L. Molecular Framework Analysis of the Generated Database GDB-13s. *J. Chem. Inf. Model.* **2023**, *63*, 484–492.
51. Flanagan, M. E.; Blumenkopf, T. A.; Brissette, W. H.; Brown, M. F.; Casavant, J. M.; Shang-Poa, C.; Doty, J. L.; Elliott, E. A.; Fisher, M. B.; Hines, M.; Kent, C.; Kudlacz, E. M.; Lillie, B. M.; Magnuson, K. S.; McCurdy, S. P.; Munchhof, M. J.; Perry, B. D.; Sawyer, P. S.; Strelevitz, T. J.; Subramanyam, C.; Sun, J.; Whipple, D. A.; Changelian, P. S. Discovery of CP-690,550: a potent and selective Janus kinase (JAK) inhibitor for the treatment of autoimmune diseases and organ transplant rejection. *J. Med. Chem*. **2010**, *53*, 8468−8484.
52. Rojas, J. J.; Bull, J. A. Oxetanes in Drug Discovery Campaigns. *J. Med. Chem.* **2023**, *66*, 12697–12709.
53. a) Tautermann, C. S.; Borghardt, J. M.; Pfau, R.; Zentgraf, M.; Weskamp, N.; Sauer, A. Towards holistic Compound Quality Scores: Extending ligand efficiency indices with compound pharmacokinetic characteristics. *Drug Discov. Today* **2023,** *28*, 103758. b) Beckers, M.; Fechner, N.; Stiefl, N. 25 Years of Small-Molecule Optimization at Novartis: A Retrospective Analysis of Chemical Series Evolution. *J. Chem. Inf. Model.* **2022**, *62*, 6002−6021.
54. a) Chakravorty, S. J.; Chan, J.; Greenwood, M. N.; Popa-Burke, I.; Remlinger, K. S.; Pickett, S. D.; Green, D. V. S.; Fillmore, M. C.; Dean, T. W.; Luengo, J. I.; Macarron, R. Nuisance Compounds, PAINS Filters, and Dark Chemical Matter in the GSK HTS Collection. *SLAS Discov.* **2018,** *23*, 532-545. b) Wassermann, A. M.; Lounkine, E.; Hoepfner, D.; Le Goff, G.; King, F. J.; Studer, C.; Peltier, J. M.;

Grippo, M. L.; Prindle, V.; Tao, J.; Schuffenhauer, A.; Wallace, I. M.; Chen, S.; Krastel, P.; Cobos-Correa, A.; Parker, C. N.; Davies, J. W.; Glick, M. Dark Chemical Matter as a Promising Starting Point for Drug Lead Discovery. *Nat. Chem. Biol*. **2015**, *11*, 958−966.

55. Waring, M. J.; Arrowsmith, J.; Leach, A. R.; Leeson, P. D.; Mandrell, S.; Owen, R. M.; Pairaudeau, G.; Pennie, W. D.; Pickett, S. D.; Wang, J.; Wallace, O.; Weir, A. An Analysis of the Attrition of Drug Candidates from Four Major Pharmaceutical Companies. *Nat. Rev. Drug Discov*. **2015**, *14*, 475−486
56. Raju, T. N. K. The Nobel Chronicles. *Lancet* **2000**, *355*, 1022.
57. Shenvi, R. A. Natural Product Synthesis in the 21st Century: Beyond the Mountain Top. *ACS Cent. Sci.* 2024. https://doi.org/10.1021/acscentsci.3c01518.

.