JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

Strategy to Target the Substrate Binding site of SET Domain Protein Methyltransferases

Kong T. Nguyen,[†] Fengling Li,[†] Gennadiy Poda,[‡] David Smil,[†] Masoud Vedadi,[†] and Matthieu Schapira*[,†,§]

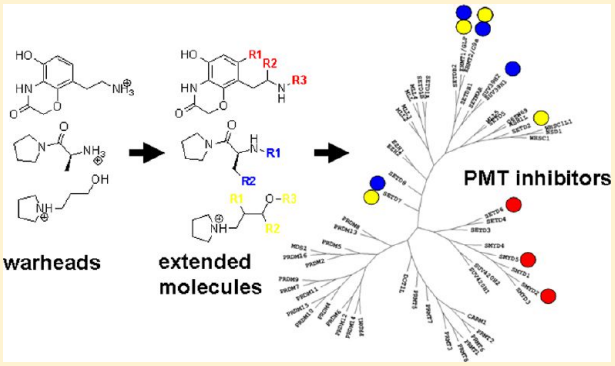
[†]Structural Genomics Consortium, University of Toronto, MaRS Centre, South Tower, seventh floor, 101 College Street, Toronto, Ontario, M5G 1L7, Canada

[‡]Medicinal Chemistry Platform, Ontario Institute for Cancer Research, 101 College Street, MaRS Centre, South Tower, Toronto, Ontario, M5G 0A3, Canada

[§]Department of Pharmacology and Toxicology, University of Toronto, Medical Sciences Building, 1 King's College Circle, Toronto, Ontario, M5S 1A8, Canada

Supporting Information

ABSTRACT: Protein methyltransferases (PMTs) are a novel gene family of therapeutic relevance involved in chromatin-mediated signaling and other biological mechanisms. Most PMTs are organized around the structurally conserved SET domain that catalyzes the methylation of a substrate lysine. A few potent chemical inhibitors compete with the protein substrate, and all are anchored in the channel recruiting the methyl-accepting lysine. We propose a novel strategy to design focused chemical libraries targeting the substrate binding site, where a limited number of warheads each occupying the lysine-channel of multiple enzymes would be decorated by different substituents. A variety of sequence and structure-based approaches used to analyze the diversity of the lysine channel of SET domain PMTs support the relevance of this strategy. We show that chemical fragments derived from published inhibitors are valid warheads that can be used in the design of novel focused libraries targeting other PMTs.



INTRODUCTION

Protein methyltransferases (PMTs) are a family of enzymes that transfer a methyl group from the cofactor S-adenosylmethionine (SAM) to a lysine or arginine side-chain of histone proteins and other substrates. These enzymes are involved in chromatin-mediated control of gene expression as well as other signaling pathways and are emerging as a novel target class for drug discovery.[1,2] While protein arginine methyltransferases (PRMTs) are organized around a canonical SAM-MT fold shared with small-molecule and DNA methyltransferases,[3] the 50 human protein lysine methyltransferases (PKMTs) are defined by a unique, structurally conserved SET domain.[4] Standing as exceptions, DOT1L, CAMKMT, and METTL21D are the only known lysine PMTs without a SET domain.[5,6]

The structural and chemical coverage of SET domain PMTs is progressing rapidly,[7−9] and in the past two years, the first enzyme−inhibitor structures in complex with PMTs were released in the PDB. While both substrate and cofactor binding sites are expected to be druggable,[7,10] to date, most cocrystallized potent and selective SET domain PMT inhibitors released in the Protein Data Bank (www.rcsb.org) are competing with the substrate lysine. Representative protein−ligand complexes include G9a-UNC0638[11] (PDB 3RJW), SMYD2-AZ505[12] (PDB 3S7B), and SETD7-inhibitor (PDB 4E47).

A common feature of these interactions is that a portion of the inhibitor anchors deeply into the substrate lysine binding channel, significantly contributing to binding, while other moieties of the compound occupy more exposed and structurally variable areas of the peptide binding groove (Figure 1A). This dramatic contribution is exemplified by the difference in activity between compound 2b (G9a IC_{50} = 330 nM) and UNC0224 (G9a IC_{50} = 15 nM), two G9a inhibitors that differ only in the presence of an alkylamine group in the most potent compound that extends into the lysine channel[13] (Figure 1B). Since the lysine-binding channel is shared by all SET domain methyltransferases, this raises the possibility that one or a limited number of chemical scaffolds may be used as warhead(s) (chemical moiety able to anchor a ligand to its protein target) that would occupy the lysine channels of multiple enzymes. Decorating this warhead with diverse substituents would mediate selectivity. Obvious candidates for such putative warheads are the chemical scaffolds occupying

Received: December 12, 2012
Published: February 15, 2013

   dx.doi.org/10.1021/ci300596x | J. Chem. Inf. Model. 2013, 53, 681−691

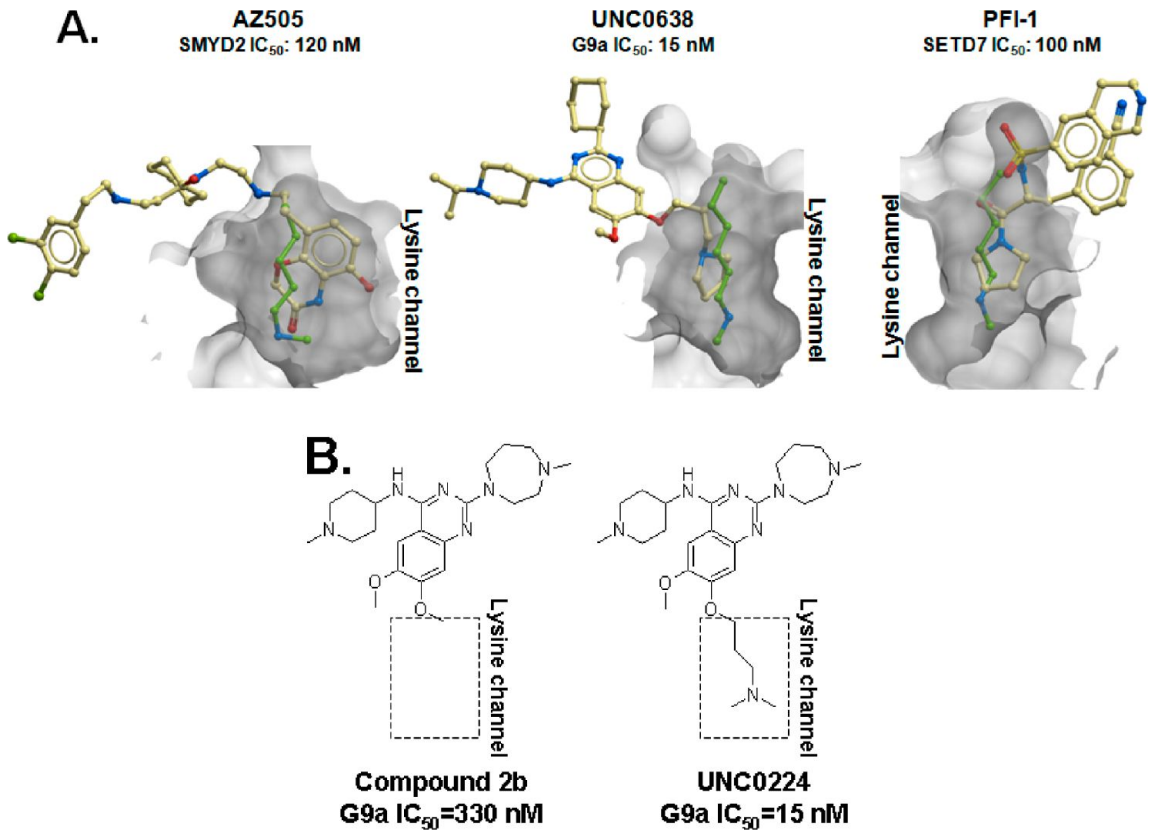


Figure 1. (A) Inhibitors competing with substrate peptides are deeply anchored in the substrate lysine binding channel (gray mesh). Co-crystallized methyl-lysine is shown in green, inhibitors are in cream. SMYD2 (PDB 3S7B), G9a (PDB 3RJW), SETD7 (PDB 4E47). (B) Addition of the alkylamine moiety that occupies the lysine channel of G9a significantly improves the activity of the inhibitor.[13]

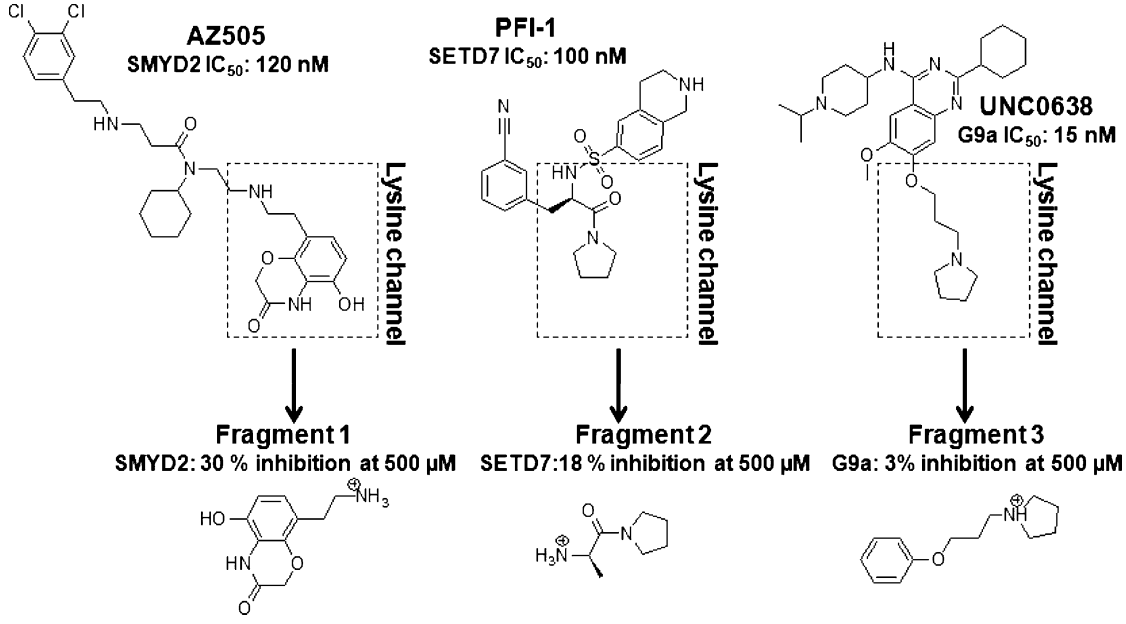


Figure 2. Biological activity of putative warheads derived from known potent inhibitors. The putative warheads are very weakly active or inactive against the enzymes with which they were cocrystallized in the context of larger molecules.

lysine channels in currently available structures (Figure 1A). We are not implying here that a fragment-based approach would necessarily identify compounds occupying the lysine channel of SET domain PMTs, as the entrance to the substrate binding groove can undergo significant conformational rearrangements upon substrate binding,[7] and it is unlikely that small molecule fragments would induce such rearrangements. Rather, we suggest a strategy where small combinatorial libraries of lead-like compounds focused on a limited set of decorated warheads, derived preferably from existing inhibitors, could be enriched in hits against enzymes with structurally related lysine channels.

dx.doi.org/10.1021/ci300596x | J. Chem. Inf. Model. 2013, 53, 681−691

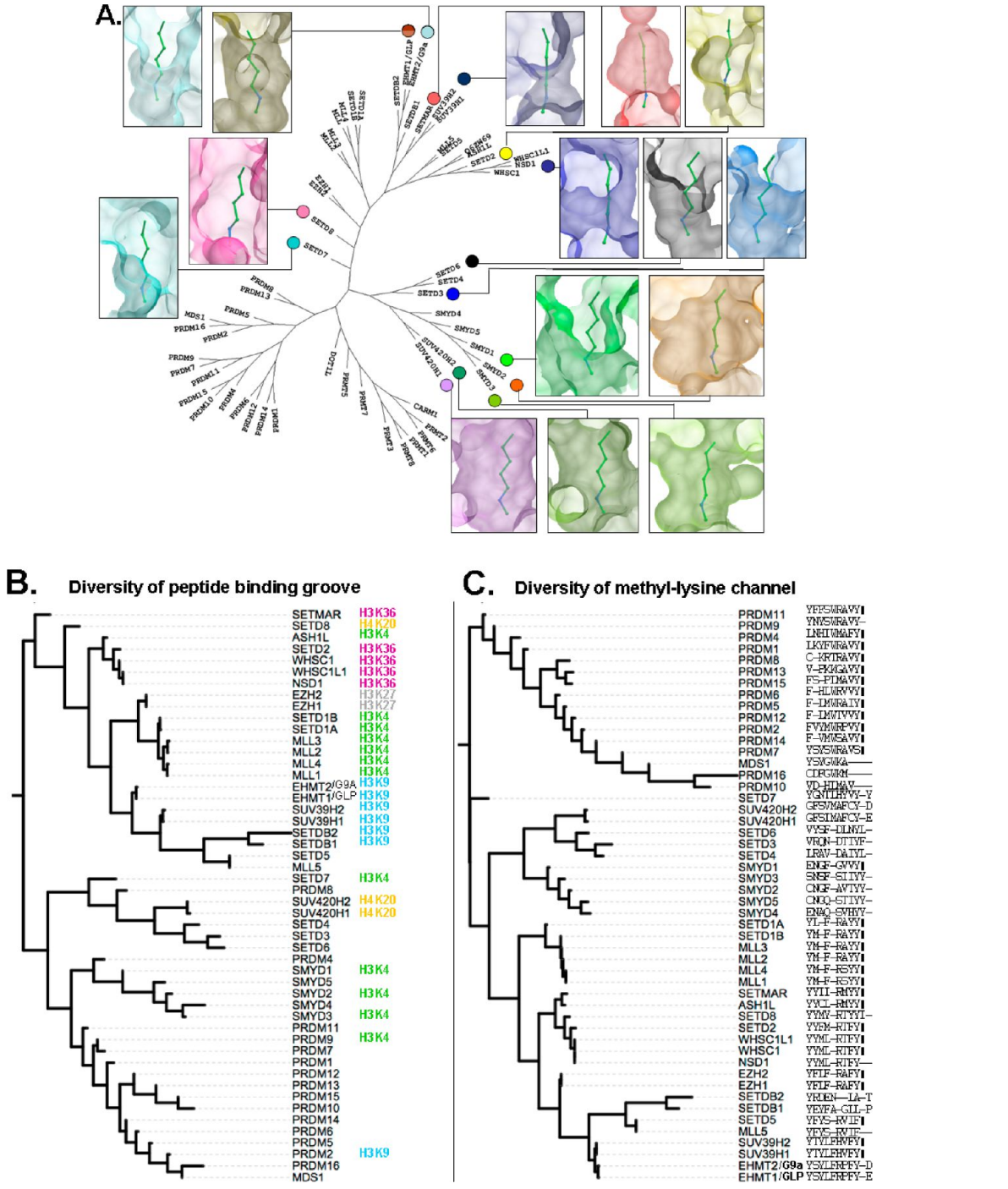


Figure 3. (A) Phylogenetic tree based on the catalytic domains of protein methyl-transferases (PMTs). Selected structures used in this study are highlighted with the shapes of their lysine channels. (B) Sequence-based clustering of the peptide-binding groove. (C) Sequence-based clustering of the methyl-lysine channels.

To explore the feasibility of this strategy, and predict the selectivity range of structurally validated scaffolds extracted from cocrystallized inhibitors, we systematically analyzed the diversity of the lysine channel of SET domain PMTs using sequence alignments, structure-based approaches and chemogenomic methods.

RESULTS

Warheads Isolated from Current Inhibitors Are Not Sufficiently Active on Their Own. The most direct way to test whether the chemical scaffolds occupying the lysine channels in available structures can act as warheads against other lysine PMTs is to test their inhibitory activity experimentally. To interrogate the validity of this approach, we first tested whether these warheads were sufficiently active against the enzymes with which they were cocrystallized in the context of a larger molecule (Figure 2). We found that the benzooxazinone-containing warhead from AZ505 (a 120 nM inhibitor of SMYD2) occupying the lysine channel of SMYD2 (PDB 3S7B), was very weakly active against the enzyme (30% ± 5% inhibition at 500 μM). The pyrrolidine warhead occupying the lysine channel of SETD7 (PDB 4E47), extracted

dx.doi.org/10.1021/ci300596x | J. Chem. Inf. Model. 2013, 53, 681−691

from a 100 nM inhibitor was inactive against SETD7, and the alkyl pyrrolidine chain occupying the lysine channel of G9a (PDB 3RJW) from the 15 nM inhibitor UNC0638 was inactive against G9a (Figure 2, and Supporting Information Figure SI-1). This result should not come as a surprise. First, it is known that fragments isolated from larger inhibitors do not always recapitulate the binding pose observed in the context of the parent molecule.[14] Second, binding of substrate peptides induce structural rearrangements of post-SET domain residues that, in the absence of substrate, can occlude the entrance to the lysine channel, as illustrated by the side-chain of Y337 in the apo structure of SETD7 (Supporting Information Figure SI-2), or alter its electrostatics, as does R1214 of GLP (Supporting Information Figure SI-2). Substrate-induced conformational rearrangements are partly recapitulated in enzyme−inhibitor structures, but fragments may be unable to induce similar fitting motion. Protein rearrangement cannot play a role in the poor activity of the SMYD2-targeted fragment: the lysine channel of SMYD2 seems to be unchanged between the apo and substrate bound structures (PDB 3TG4 and 3TG5, respectively). We note that the lysine channel is occupied by two glycerol molecules in the absence of substrate, which may compete with fragment binding (Supporting Information Figure SI-2).

These results do not necessarily indicate that the lysine channel of PMTs is not suitable for fragment screening. Indeed, fragments that exploit more efficiently this site may be detectable via screening. However, the absence of activity observed here suggests that it would be challenging to define experimentally the selectivity profile of warheads extracted from currently known inhibitors.

Amino Acid Composition of the Lysine Channel Is Diverse. To investigate the expected selectivity profile of warheads, we first clustered SET domain PMTs based on the composition of amino acids that contribute to binding site formation. A multiple sequence alignment of the entire catalytic domains was generated (see the Experimental Section for details), and residues within 4 Å of the substrate lysine side-chain in the ternary complexes of GLP, SETD7, SETD8, and SMYD2 (PDB 2RFI, 1O9S, 1ZKK, and 3TG5, respectively) were mapped on the alignment. A subalignment representing the lysine channel was generated by extracting the corresponding positions from the parent alignment, and a pseudoevolutionary clustering was produced accordingly, that reflects the similarity of amino-acids forming the lysine channel (Figure 3C). The first observation is that, even though the lysine channel is sharing the same substrate across all enzymes, and has known structural similarities (including the presence of a catalytic tyrosine, and additional aromatic residues that can contribute to methylation specificity), it is composed of a surprisingly diverse array of residues from one protein to another. It is therefore unlikely that a single fragment would optimally satisfy structural variations across all enzymes. It is nevertheless possible that a limited set of fragments may serve as warheads against numerous enzymes.

To complement this analysis, we also extracted from the parent multiple sequence alignment all positions within 4 Å of the entire cocrystallized substrate peptides from the same GLP, SETD7, SETD8, and SMYD2 structures. We note that the similarity profiles based on the sequence of the entire catalytic domain (Figure 3A), the substrate peptide binding groove (Figure 3B), and the lysine channel (Figure 3C) are very similar and tend to cluster enzymes that share histone substrates (highlighted in Figure 3B). Taken together, these results indicate that the lysine channel of SET domain PMTs is composed of a variable array of amino acids, that the variability of amino acids reflects with surprising fidelity the variability of the substrate peptide binding groove and of the entire catalytic domain, and that a single chemical scaffold will probably not fit optimally in all lysine channels.

Clusters of Structurally Related Lysine Channels Can Be Defined. Variability in the nature of side-chains composing a binding site does not necessarily reflect accurately variability in structure. We therefore used the atomic-property field (APF) method implemented in ICM[15] to compare the spatial arrangement and property of atoms forming the lysine channel of fifteen human SET domain PMTs for which a high resolution structure is available (structures highlighted in Figure 3A). Since substrate or inhibitor-bound structures are available for only six of these enzymes (G9a 3K5K, GLP 3HNA, SETD6 3QXY, SETD7 1O9S, SETD8 1ZKK, and SMYD2 3S7D), we made the decision to use cofactor-bound structures for this analysis, regardless of the presence of substrate or inhibitor. Indeed, available data indicate that the most drastic conformational rearrangement of the lysine channel takes place upon cofactor binding;[7] once the cofactor is bound, the lysine channel seems to adopt a conformation very close to the one observed in substrate-bound ternary complex. The entrance to the lysine channel is however sometimes occluded by post-SET residues projecting into the substrate binding groove.[7] When the case, we removed these residues before using the structure (see the Experimental Section and Supporting Information Table SI-1 for details). When available, we included both substrate-bound and substrate-free structures (which we later refer to as "apo", even though the cocrystallized cofactor is present) in the analysis.

This structure-based approach clearly defines four groups of enzymes sharing overall similar lysine channels (Figure 4A): (i) G9a and GLP, (ii) SMYD2, SMYD3, and SETD6, (iii) SUV420H1 and SUV420H2, (iv) SMYD1 and SETD3. Importantly, we note that apo structures of SETD7, GLP, and SMYD2 cluster in the same groups as the corresponding substrate-bound structures, which supports the inclusion of apo structures in the analysis. While structural similarities between G9a and GLP or SUV420H1 and SUV420H2 are obvious from the sequence based clustering, the composition of the other two groups is not trivial. To test the validity of the atom-based APF method, we verified visually that lysine channels were structurally similar within clustered enzymes and structurally dissimilar between nonclustered enzymes, as illustrated for the clustered pair SETD6/SMYD2 (APF distance = 0.89) and the nonclustered pair SETD7/SMYD3 (APF distance = 2.85; Figure 4B). The fact that SETD6 and SMYD2 are not close neighbors in the dendrogram derived from the similarity of residues lining the lysine channel (Figure 3C) can result from at least two factors: (1) multiple alignments of large numbers of diverse sequences are prone to inaccuracies, even when structural information is used to generate the alignment, as was the case here and (2) the set of positions selected to represent the lysine channel is derived from a limited number of structures, and is unlikely to be conserved across all enzymes, with the risk of missing important positions or including irrelevant ones.

These results show that the overall distribution and nature of atoms forming the lysine channel can be used to cluster SET domain PMTs in groups of enzymes, such as SMYD2, SMYD3,

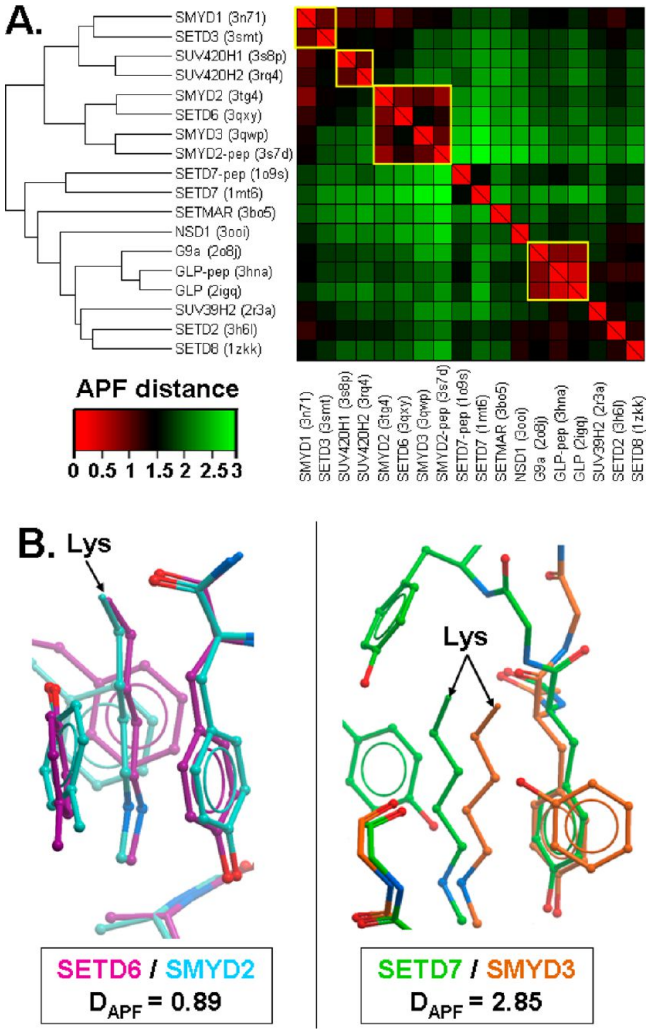


Figure 4. (A) Structural clustering of human methyltransferases lysine sites using ICM APF method.[15] Structures separated by small APF distances are similar. Larger APF distances are associated with increased structural divergence. (B) SMYD2 (PDB 3TG4 and SETD6 (PDB 3QXY) are structurally similar, resulting in a low APF distance of 0.89. SETD7 (PDB 1O9S) and SMYD3 (PDB 3QWP) are structurally dissimilar and separated by a large APF distance of 2.85.

and SETD6, more likely to be recognized by unique lysine-channel targeting warheads and that this clustering method is more accurate than a sequence-based approach.

Spatial Distribution of Interaction Hot Spots Can Be Used to Compare Ligand Interaction Potentials. It is commonly acknowledged that biomolecular binding events are generally driven by a limited set of key contacts at interaction hot spots, which are positions of high interaction potentials.[16] A drawback to the structure-based method described above is that it treats equally all atoms of the lysine channel, while conservation of atoms with strong interaction potentials, which are overwhelmingly contributing to the binding energy, is more important than conservation of atoms that are less likely to contribute significantly to binding. GRID is a common computational tool used to map the interaction potential along binding pockets.[17] While it is very useful to characterize a single binding site, we find the mesh representation of energy potentials difficult to superimpose and compare from one site to another. We therefore applied a novel method, inspired by recent work from Vajda and colleagues,[18,19] to map hot spots. Briefly, a collection of 1589 chemically diverse fragments was docked with Glide XP[20] (Schrodinger, NY) to the lysine channels structures, and ICM was used to count the number of hydrogen bonds formed by all ligands at each atomic position of the binding pocket, which represents the strength of the interaction potential at this position. The fragment docking and scoring method was first tested for its ability to retrieve structurally validated fragments (i.e., fragments cocrystallized in the context of larger inhibitors) from the 1589 diverse fragment set (see the Experimental Section and Supporting Information Table SI-2 for details). We decided here to focus exclusively on polar hot spots, as selectivity profiles of ligands are most influenced by polar interactions. We verified that this method recapitulated accurately results produced by the commonly accepted method GRID, implemented in MOE (CCG, Canada) (Supporting Information Figure SI-3). Since this method is based on ligand docking, which, unlike APF comparison of binding site structures, can be very sensitive to moderate motion of a single side-chain, we focused this analysis on structures where the lysine channel is occupied by relevant chemical matter, which is expected to induce optimal positioning of surrounding side-chains. This limited the analysis to substrate-bound structures of SETD7, GLP, SETD8, SETD6, and SMYD2 and the inhibitor-bound structure of G9a.

Interestingly, two hot spots are observed with variable strengths in all structures (Figure 5). The first one is at a position corresponding to the backbone carbonyl of T266 in SETD7. This position forms a hydrogen bond with the amide nitrogen of substrate lysine in all available substrate-bound structures,[7] and is also exploited by UNC0638 and AZ505, two potent inhibitors of G9a and SMYD2, respectively,[11,12] which supports its prediction as a position of strong interaction potential. This hot spot is particularly strong in SETD7, GLP, and SETD6, and is likely to contribute heavily to binding of warheads targeting these three enzymes. The second conserved hot spot is located at a position corresponding to the carbonyl of G264 in SETD7. This carbonyl group is believed to enhance the nucleophilicity of SAM's departing methyl group during catalysis,[21] which explains its presence in all structures analyzed. Displaying hot spots on the superimposed structures clearly highlights shared positions of strong interaction potentials (Figure 5, bottom right panel). However, the strength of the potentials varies from one enzyme to another. We also note that the hot spot profiles of SMYD2 and SETD6 are very close and composed of two hydrogen-bond acceptors at the backbone carbonyl of S224, A222 in SETD6 and G183, C181 in SMYD2 and a hydrogen bond donor at the side-chain phenolic hydroxyl of Y297 in SETD6 and Y258 in SMYD2.

These results reveal that the distribution and strength of interaction hot spots along the lysine binding channel varies from one enzyme to the other and are conserved between specific groups of enzymes, which are more likely to be targeted by identical warheads.

Fragment Docking Profiles Indicate Unsuspected Similarities in the Structural Chemistry of Lysine Channels. While the spatial distribution of hot spots along a binding site dictates in part the nature of the ligands that can occupy the site, other factors, such as the shape of the pocket, are also important. A key property of diverse fragment libraries is that they sample more efficiently the chemical space than libraries composed of larger molecules.[22] A corollary is that screening a fragment library against a specific binding pocket interrogates more efficiently the structural features of this

dx.doi.org/10.1021/ci300596x | J. Chem. Inf. Model. 2013, 53, 681−691

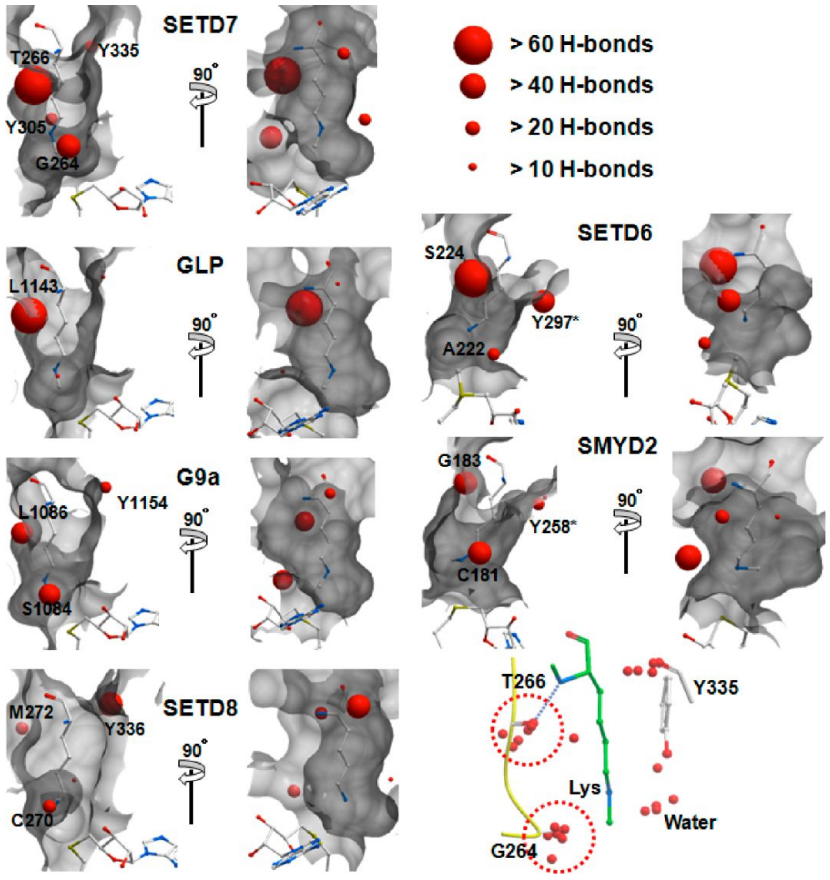


Figure 5. Hot spots for the formation of hydrogen bonds identified by docking a fragment library. The strength of the interaction potential at any atom of the binding site is reflected by the cumulative number of hydrogen bonds formed by the docked fragments. The location and strength of hot spots vary from one structure to the other, but two are conserved, located at the main-chain carbonyl oxygens of G264 and T266 in the SETD7 lysine site (bottom right).

pocket. We therefore hypothesized that the collection of fragments that fit well into a specific lysine channel could represent a chemical profile of the channel and that pockets sharing similar chemical profiles would be more likely to be recognized by the same ligands. To generate such chemical profiles we selected for each lysine channel the top 100 Glide XP scoring fragment binding poses out of our 1589 fragment collection. All poses had to form a hydrogen bond with the conserved hot spot previously identified and exploited by cocrystallized G9a and SMYD2 inhibitors (corresponding to carbonyl oxygen of T266 in SETD7). The number of poses shared between two enzymes (implying identical ligand, with a binding pose RMSD lower than 1.5 Å upon superimposition of the receptors) was used as a measure of the distance separating the chemical profiles of their two binding sites. The resulting array of distances was transformed into a distance matrix (Figure 6).

First, we observe that the results from previous methods are partly recapitulated: SETD6 and SMYD2 are clustered closely, indicating similar chemical profiles, and increased chances of sharing warheads, as was indicated by directly comparing their structures (Figure 4), or comparing the distribution of their hot spots (Figure 5). Similarly G9a and GLP are grouped in the same cluster, which was expected considering their high sequence and structure similarity. More surprisingly, this approach clusters SETD7 with GLP and G9a (Figure 6), which is not found by comparing their structures (Figure 4). This is probably resulting from (1) the high similarity in the

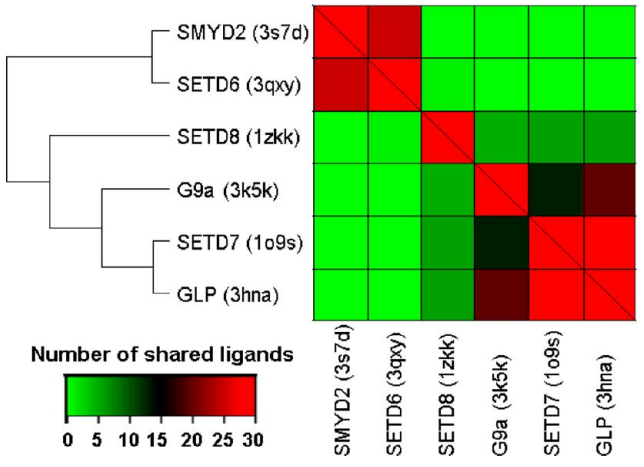


Figure 6. Chemogenomic clustering of the lysine-binding sites of selected PMTs. The degree of similarity between two lysine sites is measured by the total number of shared ligands. A shared ligand is a docked fragment that scores well against two lysine sites with similar poses (RMSD ≤ 1.5 Å).

overall geometry of their lysine channels, as illustrated by mesh representation of their van der Waals boundaries (Figure 5), this, in spite of the divergence in nature and position of most atoms constituting the channels, and (2) the strong hot spot shared by the three pockets at a position corresponding to the carbonyl oxygen of T266 in SETD7. SETD8 has a much more

dx.doi.org/10.1021/ci300596x | J. Chem. Inf. Model. 2013, 53, 681−691

open lysine channel, and a weaker hot spot at the corresponding M272, and the pockets of SETD6 and SMYD2 have very different shapes. Unlike structure comparison, fragment docking profiles successfully captured the similarity in the structural chemistry of the GLP and SETD7 pockets, defined as the shape of the pockets and positions of high interaction potentials.

To support this result, we inspected the shared binding poses of top scoring ligands docked in the lysine channels of clustered enzymes. Some fragments are docked in the same orientation in the lysine channel of GLP and SETD7 and make the same set of interactions with surrounding interaction hot spots, as well as a water molecule conserved in all structures (Figure 7).

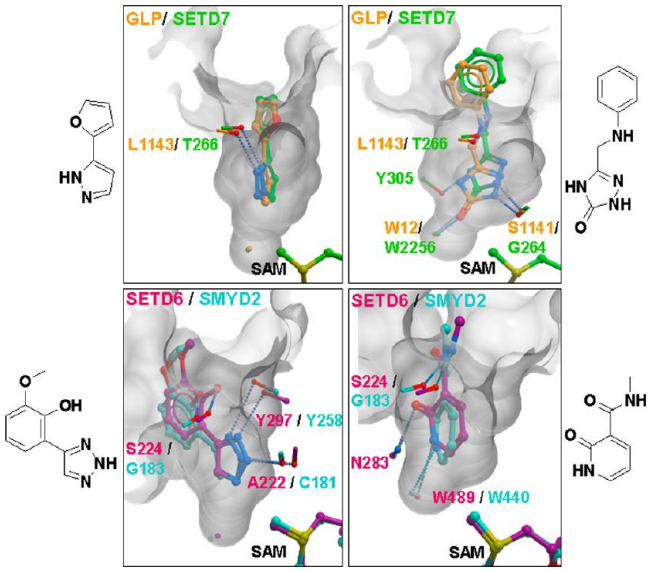


Figure 7. Examples of docked ligands shared by two different lysine channels. The ligands show similar docking poses and make the same hydrogen bonding network in the lysine sites of GLP and SETD7 (top) and SETD6 and SMYD2 (bottom).

Similarly, some docked fragments adopt a conserved pose and make conserved hydrogen bonds with surrounding atoms in SETD6 and SMYD2. These structures are not experimental, but computational models, and not all of these fragments may bind as predicted by docking, but the putative complexes do illustrate that the shape and chemical environment of the lysine channels for these enzyme pairs can be exploited by the same set of ligands.

We were surprised to observe that a limited number of docked fragments were selected against up to five PMTs, making a conserved set of interactions with surrounding atoms (Supporting Information Figure SI-4). While the number of virtual hits shared by SETD7 and GLP or SMYD2 and SETD6 was high (up to 30), fragments selected against four or five targets were very rare (only 4) and may be an artifact due to the high rate of false positives that is inherent to any virtual screening exercise. For instance, no experimental data to date has confirmed that a six member ring can fit in the lysine channel of SETD7, G9a, or GLP, as is predicted in the example shown.

Together, these results suggest that a unique warhead could target optimally SETD7, GLP, and G9a, and another SETD6 and SMYD2.

DISCUSSION

On the basis of available structural and biochemical data showing that known potent inhibitors of SET domain PMTs that compete with the substrate are deeply anchored in the lysine channel, and based on experimental data presented here showing that structurally validated fragments occupying lysine channels are not necessarily active in biochemical assays, we propose a strategy for the design of chemical libraries for this protein family where a limited number of warheads targeting the lysine channel of multiple enzymes are decorated with diverse substituents that mediate selective inhibition. We analyzed the structural diversity of the lysine channel across available structures by successively applying multiple methods with increased resolution, starting with a less accurate

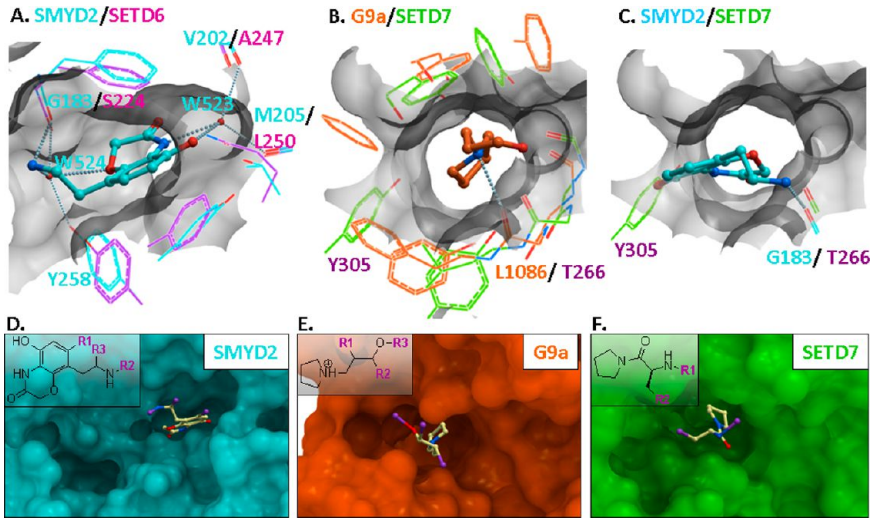


Figure 8. Pairwise superpositions of lysine channels indicate how scaffolds cocrystallized with one enzyme fit into another. (A and C) Fragment extracted from the cocrystallized SMYD2 inhibitor AZ505 fitting well in the lysine channel of SETD6, but not in SETD7. (B) Pyrrolidine moiety of G9a inhibitor UNC0638 fitting well in the lysine channel of SETD7. (D–F) Binding poses of fragments derived from the cocrystal structures of active inhibitors indicating open positions where warheads can be functionalized to achieve diverse target selectivity profiles: (blue) nitrogen, (red) oxygen, (magenta) R-group.

dx.doi.org/10.1021/ci300596x | J. Chem. Inf. Model. 2013, 53, 681–691

sequence-based approach that covers all PMTs, continuing with a more precise APF method that requires at least apo structures of the enzymes, and ending with a more sensitive chemogenomic approach that is limited to proteins for which a ternary complex in the presence of cofactor and substrate is available. Our results suggest that two distinct warheads are sufficient to target optimally the lysine binding sites of SMYD2 and SETD6 for the first and that of GLP, G9a, and SETD7 for the second. In support of this model, we note that superimposing a structure of SETD6 to that of SMYD2 in complex with the inhibitor AZ505 places the fragment of the compound occupying the lysine channel of SMYD2 in a seemingly perfect orientation in the lysine channel of SETD6 (Figure 8). Similarly, the pyrrolidine-based fragment occupying the lysine channel of G9a fits favorably in the lysine channel of SETD7 when superimposing SETD7 to the G9a cocrystal structure, even though surrounding residues are not well conserved. On the other end, the fragment extracted from the SMYD2 complex does not fit in SETD7, due to a clash with Y305.

The chemical profiles that we use to cluster targets are based on fragment docking, and one may be tempted to directly test the docked fragments, rather than use the docking results as a measure of structural similarity. We believe that this effort may not lead to the desired outcome for two reasons. First, preliminary data shown in this work indicates that fragments cocrystallized in the lysine channels of G9a, SETD7, and SMYD2 are essentially inactive against G9a, SETD7, and SMYD2 when extracted from the larger potent inhibitor. It is therefore likely that potentially valid fragments selected by virtual screening would be inactive: chemical synthesis of larger analogues would be necessary to test their validity. Second, false positive rates of 90% or higher are common in virtual screening, sometimes because of minor van der Waals clashes that were overlooked by the scoring function, or because of conformational strains associated with the docked pose. Nevertheless, even if a docked compound may actually not quite fit in a pocket, or if the docked compound is strained, the network of hydrogen bonds and overall shape captured by the docked pose generally remain valid and can be used to define the structural chemistry of the binding site, as was done here. For this reason, we believe that the best approach to design a diverse library of lead-like PMT inhibitors is to decorate warheads that have been experimentally validated. As the number of structures crystallized in complex with substrate or inhibitor increases, our clustering of SET domain PMTs will gain in completion and the number of warheads necessary to cover the entire family will become clear.

CONCLUSION

Protein methyltransferases have recently emerged as an attractive drug target family, and only a small number of potent and selective chemical inhibitors have been reported.[2] New strategies have to be developed to design chemical libraries of potent inhibitors focused on this protein family. An approach is to design compounds that occupy the cofactor binding site shared by all PMTs. A complementary strategy presented here is focused on the substrate-binding site. This method relies heavily on the availability of high resolution structures and a limited set of validated inhibitors that can be used as seeds for target hopping. We believe that we are rapidly reaching a position where this strategy can be applied efficiently.

EXPERIMENTAL SECTION

Methyltransferase Activity Assay. Methyltransferase activity assays for G9a, SETD7, and SMYD2 were performed by monitoring the transfer of tritium-labeled methyl groups to peptide substrates using scintillation proximity assay (SPA). A peptide corresponding to the first 25 residues of histone H3 (1−25) was used as a substrate for G9a and SETD7. For SMYD2 a peptide corresponding to residues 361−380 of p53 was used as a substrate. For G9a, the assay mixture contained 5 nM G9a, 1 μM biotinylated peptide H3 (1−25), and 8 μM SAM in 25 mM potassium phosphate pH 8.0, 1 mM EDTA, 2 mM $MgCl_2$, and 0.01% Triton X-100. The SETD7 reaction mixture contained 20 nM SETD7, 2 μM biotinylated peptides H3 (1−25), 2 μM SAM in 20 mM Tris pH 8.0, 5 mM DTT, and 0.01% Triton X-100. SMYD2 assay mixture contained 30 nM SMYD2, 3 μM biotinylated p53 (361−380), 0.5 μM SAM in 50 mM Tris-pH 9.0, 2 mM DTT, and 0.02% Tween 20.

IC_{50} values were determined using fragment concentrations ranging from 500 nM to 500 μM, and control compound concentrations ranging from 5 nM to 5 μM. To stop the enzymatic reactions, 7.5 M guanidine hydrochloride was added, followed by 180 μL of buffer (20 mM Tris, pH 8.0), mixed and then transferred to a 96-well FlashPlate (cat. no. SMP103; Perkin-Elmer; www.perkinelmer.com). After mixing, the reaction mixtures in the Flash plate were incubated for 1 h, and the CPM counts were measured using the Topcount plate reader (Perkin-Elmer, www.perkinelmer.com). The CPM counts in the absence of compound for each data set were defined as 100% activity. In the absence of the enzyme, the CPM counts in each data set were defined as background (0%). The IC_{50} values were determined using SigmaPlot software.

Purification of Commercial Compounds. Purity determination was conducted by UV absorbance at 254 nm during tandem liquid chromatography/mass spectrometry (LCMS) using a Waters Acquity separations module. Identity was determined via low-resolution mass spectra (LRMS) acquired in positive ion mode using a Waters Acquity SQD mass spectrometer (electrospray ionization source) fitted with a PDA detector. Mobile phase A consisted of 0.01% formic acid in water, while mobile phase B consisted of 0.01% formic acid in acetonitrile. The gradient ran from 5% to 95% mobile phase B over 5 min at 0.5 mL/min. An Acquity CSH C18, 1.7 μm, 2.1 × 50 mm column was used with column temperature maintained at 25 °C. The sample solution injection volume was 5 μL. The identity of all compounds analyzed by this method was confirmed, and recorded purities were ≥95%.

Phylogenetic Tree of PMT. The tree was taken from the ChromoHub[23] Web site and decorated with surface meshes of 15 selected lysine channels (Figure 3A). The surface meshes were generated using ICM (version 3.7−2c; MolSoft: San Diego, CA, 2012).

Multiple Sequence Alignment. Whole substrate peptide-groove: all residues within 4 Å of the peptides in four peptide-bound structures: SETD8 (PDB 1ZKK); SETD7 (PDB 1O9S); SMYD2 (PDB 3TG5); and GLP (PDB 2RFI) were selected using the ICM selection expression *as_graph = Res(Sphere-(a_1zkk.pep a_1zkk.1//!n,c,o,h* 4.)) | Res(Sphere(a_1o9s.pep a_1o9s.1//!n,c,o,h* 4.)) | Res(Sphere(a_2rfi.pep a_2rfi.1//!n,c,o,h* 4.)) | Res(Sphere(a_3tg5.pep a_3tg5.1//!n,c,o,h* 4.))*

This residue selection was propagated to corresponding positions in the ChromoHub multiple sequence alignment. The

Newick string generated from the alignment was used as input to generate images in iTol tree of life[24] (Figure 3B).

Lysine-site: The following ICM selection expression *Res-(Sphere(a_pep./substrate_lysine a_1//!n,c,o,h* 4.))* focusing on residues lining the lysine site was used to extract the relevant positions from from for the ChromoHub alignment. A Newick string of the corresponding alignment was used as input to generate images in iTol tree of life[24] (Figure 3C).

Target Clustering Based on APF Distances. A set of 15 targets marked with circle as shown in Figure 3 were used. In case of GLP, SETD7, and SMYD2, both apo and peptide-bound (holo) structures were available and were included into this analysis, resulting in a total of 18 structures. All SAH/SAM and crystal water molecules were removed. In seven targets including NSD1, SETD2, SETD3, SETMAR, SUV39H2, SUV420H1, and SUV420H2, autoinhibitory residues that blocked the lysine site were removed (details in Supporting Information Table SI-1). Lysine side-chains of cocrystallized substrates of SMYD2, GLP, and SETD7 were used to identify lysine channels for all structures of these targets. For other targets without a cocrystallized substrate, the side-chain of lysine stemmed from peptide substrate of GLP (PDB 3HNA) was used to define the lysine channels. ICM APF site-superposition method was used to measure the distance between any two lysine sites. First, the neighborhood within 6 Å around the inserted lysine peptide of site A and site B were extracted and then superimposed on each other. Upon reaching the optimal superposition, APF pseudoenergy or E_{APF} between site A and B were calculated. E_{APF} is always negative and its value depends on size and compositions of sites. E_{APF} values were then converted to normalized dot product-like measures with correct asymptotic behavior using the following formula:[15]

$$S_{APF} = -\tanh((E_{APF} - E_0)/\Delta_0)$$

where $E_0 = 100$ and $\Delta_0 = -250$ are suggested empiric parameters. Next, distance-like similarity measure was obtained from the dot-product-like:

$$D_{APF}(A, B) = (S_{APF}(A, A) + S_{APF}(B, B) - 2S_{APF}(A, B)$$

Finally, a distance matrix for all pairs of lysine sites was generated. The APF distances range from 0 to 2.90, with null value indicates the distance between two identical sites. The greater the APF distance, the more dissimilar the two lysine sites. The distance matrix was imported into R to generate a heat-map using average distance (UPGMA) clustering method.

Fragment Docking. *Receptors Preparation.* Six holo-structures including G9a (PDB 3K5K), GLP (PDB 3HNA), SETD6 (PDB 3QXY), SETD7 (PDB 1O9S), SETD8 (PDB 1ZKK), and SMYD2 (PDB 3S7D) with resolution less than 2.5 Å were used in docking study. Four of these structures, namely G9a, GLP, SETD7, and SETD8, were solved with a bound SAH which was uniformly replaced by a SAM. All six structures were superimposed via superposition of their SAMs. First, protein receptors were imported and preprocessed in MolSoft ICM by calling the "convert PDB" function which automatically adds missing side-chains if any. The preprocessed receptor structures were then loaded in Schrodinger Maestro for protein preparation and Glide docking. All default steps in Maestro Protein Preparation Wizard were followed, including bond orders assignment, removal of the original hydrogen atoms and addition of new hydrogen atoms. Protonation states were set at pH 7.4. Next, H-bond assignment and orientation of conserved water molecules were sampled exhaustively by Protassign at neutral pH. Finally, a short Impref minimization with RMSD cutoff of 0.3 Å for all heavy atoms convergence was implemented using an OPLS2005 force field.

Fragment Set Selection. A small yet highly diverse fragment set has been selected for this study. We compiled such a ligand set from the ZINC database that already contains multiple ligand subsets suitable for virtual screening. For fragment docking, we used the ZINC fragment set downloaded from the ZINC database Web site,[25] imported it into ICM MolCart and generated a set of 1589 chemically diverse fragments with molecular weight evenly distributed from 120 to 223 Da. Parts of known inhibitors of three PMT lysine sites G9a, SETD7, and SMYD2, occupying the corresponding lysine sites provide an estimate on what size fragments can bind there. The upper cutoff of 223 comes from the molecular weight of the biggest known fragment that can still fit into the defined docking grid-box. The fragment set was then converted into 3D structures using Schrodinger's LigPrep with protonation states set at pH = 7.4 using Epik.

Glide XP docking: A cubic docking box of 12 Å × 12 Å × 12 Å containing an internal box of 10 Å × 10 Å × 10 Å sufficiently covering the whole inserted putative lysine side-chain up to the branching point of a substrate peptide was centered on all six lysine sites. All hydroxyls of Tyr, Ser, and Thr residues located inside the docking box and directly pointing toward the lysine site were set rotatable during the Glide grid calculation. The Glide XP (version 57109; Schrödinger: New York, NY, 2012) flexible docking mode was used. At the final step of each docking run, a total of 100 docking poses per ligand were kept for postdocking minimization, from which a maximum of 10 poses per ligand were selected and saved in the output file. Docking poses in the output file were sorted based on the original Glide XP scoring function which may not be optimized for small fragments. To this end, a scoring function taking into account ligand efficiency (LE) was used to rescore the original XP output. This LE-based rescoring tool can be downloaded from the Schrodinger Script Center as a command-line script named "fragment_selector.py". This script takes a Glide pose-view as an input file, and rescores and filters docked fragment poses using the ligand efficiency metric and spatial diversity. For ligand efficiency, the Glide XP score is normalized by the number of heavy atoms. By default, natural log ligand efficiency is used. For spatial diversity, the script takes the top poses by score for each region of the active site.

Identification of Hydrogen-Bonding Hot Spots. The top 100 ranking ligands from each Glide XP docking were used as probe set to identify hydrogen-bonding hot spots on the surface of the lysine binding site. First, all hydrogen-bond donor (HBD) and hydrogen-bond acceptor (HBA) atoms within 4 Å from the lysine side chain of the substrate peptide, including the conserved water, were extracted and stored in an atomic array H. Then, each of the top 100 docked ligands was consecutively displayed in its binding site where the number of possible hydrogen bonds formed between the docked ligand and each atom in the array H were registered. These numbers could supposedly vary from 0, meaning no single H-bond can be formed with any of the top 100 docking poses, to a maximum of 100. In fact, not all top docked ligands contain polar moieties which can make hydrogen bonds with the corresponding lysine binding site. The lysine-site atoms that made at least 10 hydrogen bonds with the top 100 docked poses were labeled as hot spot atoms.

dx.doi.org/10.1021/ci300596x | J. Chem. Inf. Model. 2013, 53, 681−691

Ligand-Based Target Clustering. The top 100 docked poses of each of 6 targets, or 600 docked poses, were grouped together. Ligands that appeared in the top 100 of at least any two targets were annotated. For each of these annotated ligands, pairwise RMSD values between its two docked poses in two corresponding targets were calculated. Two targets are considered sharing the same ligand if RMSD between the two docked poses is less than or equal to 1.5 Å.

GRID Interaction Hot Spots. The GRID maps applied here are based on the work of Goodford et al.[17] and Boobbyer et al.[26] and were implemented in the Molecular Operating Environment (MOE) modeling suite (version 2011; Chemical Computing Group Inc.: Montreal, Canada, 2011). A variety of probes are available, spanning a range of different combinations of size, charge, and hydrogen bond donor and/or acceptor properties. In this work, water probes were used to find hydrogen-bonding hot spots. All the water probe GRID maps were calculated using the MOE application "Compute Interaction Potential" and displayed at a constant level of −5.5 kcal/mol.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

Effect of fragments on activity of PMTs (Figure SI-1); change in lysine channels between apo and holo structures (Figure SI-2); comparison of hydrogen-bonding hot spots identification methods (Figure SI-3); ligand that docks similarly to the five lysine channels (Figure SI-4); protein preparation (Table SI-1); results of different rescoring parameters (Table SI-2). This information is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*Phone: +1 416-978-3092. E-mail: matthieu.schapira@utoronto.ca.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We thank Eli Lilly & Co for providing some of the chemical fragments tested in this work. The Structural Genomics Consortium is a registered charity (number 1097737) that receives funds from Canadian Institutes of Health Research, Eli Lilly Canada, Genome Canada, GlaxoSmithKline, the Ontario Ministry of Economic Development and Innovation, the Novartis Research Foundation, Pfizer, Abbott, Takeda and the Wellcome Trust. Basic funding for OICR is provided by the Ontario government through the Ministry of Economic Development and Innovation, Canada.

■ ABBREVIATIONS

PMTs, protein methyltransferases; APF, atomic property field; XP, extra precision; RMSD, root-mean-square deviation; PDB, protein data bank

■ REFERENCES

(1) Copeland, R. A.; Solomon, M. E.; Richon, V. M. Protein methyltransferases as a target class for drug discovery. *Nat. Rev. Drug Discovery* **2009**, *8*, 724−732.

(2) Arrowsmith, C. H.; Bountra, C.; Fish, P. V.; Lee, K.; Schapira, M. Epigenetic protein families: a new frontier for drug discovery. *Nat. Rev. Drug Discovery* **2012**, *11*, 384−400.

(3) Martin, J. L.; McMillan, F. M. SAM (dependent) I AM: the S-adenosylmethionine-dependent methyltransferase fold. *Curr. Opin. Struct. Biol.* **2002**, *12*, 783−793.

(4) Qian, C.; Zhou, M. M. SET domain protein lysine methyltransferases: Structure, specificity and catalysis. *Cell. Mol. Life Sci.* **2006**, *63*, 2755−2763.

(5) Min, J.; Feng, Q.; Li, Z.; Zhang, Y.; Xu, R. M. Structure of the catalytic domain of human DOT1L, a non-SET domain nucleosomal histone methyltransferase. *Cell* **2003**, *112*, 711−723.

(6) Kernstock, S.; Davydova, E.; Jakobsson, M.; Moen, A.; Pettersen, S.; Maelandsmo, G. M.; Egge-Jacobsen, W.; Falnes, P. O. Lysine methylation of VCP by a member of a novel human protein methyltransferase family. *Nat. Commun.* **2012**, *3*, 1038.

(7) Schapira, M. Structural Chemistry of Human SET Domain Protein Methyltransferases. *Curr. Chem. Genomics* **2011**, *5*, 85−94.

(8) Richon, V. M.; Johnston, D.; Sneeringer, C. J.; Jin, L.; Majer, C. R.; Elliston, K.; Jerva, L. F.; Scott, M. P.; Copeland, R. A. Chemogenetic analysis of human protein methyltransferases. *Chem. Biol. Drug Des.* **2011**, *78*, 199−210.

(9) Yost, J. M.; Korboukh, I.; Liu, F.; Gao, C.; Jin, J. Targets in epigenetics: inhibiting the methyl writers of the histone code. *Curr. Chem. Genomics* **2011**, *5*, 72−84.

(10) Campagna-Slater, V.; Mok, M. W.; Nguyen, K. T.; Feher, M.; Najmanovich, R.; Schapira, M. Structural chemistry of the histone methyltransferases cofactor binding site. *J. Chem. Inf. Model.* **2011**, *51*, 612−623.

(11) Vedadi, M.; Barsyte-Lovejoy, D.; Liu, F.; Rival-Gervier, S.; Allali-Hassani, A.; Labrie, V.; Wigle, T. J.; Dimaggio, P. A.; Wasney, G. A.; Siarheyeva, A.; Dong, A.; Tempel, W.; Wang, S. C.; Chen, X.; Chau, I.; Mangano, T. J.; Huang, X. P.; Simpson, C. D.; Pattenden, S. G.; Norris, J. L.; Kireev, D. B.; Tripathy, A.; Edwards, A.; Roth, B. L.; Janzen, W. P.; Garcia, B. A.; Petronis, A.; Ellis, J.; Brown, P. J.; Frye, S. V.; Arrowsmith, C. H.; Jin, J. A chemical probe selectively inhibits G9a and GLP methyltransferase activity in cells. *Nat. Chem. Biol.* **2011**, *7*, 566−574.

(12) Ferguson, A. D.; Larsen, N. A.; Howard, T.; Pollard, H.; Green, I.; Grande, C.; Cheung, T.; Garcia-Arenas, R.; Cowen, S.; Wu, J.; Godin, R.; Chen, H.; Keen, N. Structural basis of substrate methylation and inhibition of SMYD2. *Structure* **2011**, *19*, 1262−1273.

(13) Liu, F.; Chen, X.; Allali-Hassani, A.; Quinn, A. M.; Wasney, G. A.; Dong, A.; Barsyte, D.; Kozieradzki, I.; Senisterra, G.; Chau, I.; Siarheyeva, A.; Kireev, D. B.; Jadhav, A.; Herold, J. M.; Frye, S. V.; Arrowsmith, C. H.; Brown, P. J.; Simeonov, A.; Vedadi, M.; Jin, J. Discovery of a 2,4-diamino-7-aminoalkoxyquinazoline as a potent and selective inhibitor of histone lysine methyltransferase G9a. *J. Med. Chem.* **2009**, *52*, 7950−7953.

(14) Babaoglu, K.; Shoichet, B. K. Deconstructing fragment-based inhibitor discovery. *Nat. Chem. Biol.* **2006**, *2*, 720−723.

(15) Totrov, M. Ligand binding site superposition and comparison based on Atomic Property Fields: identification of distant homologues, convergent evolution and PDB-wide clustering of binding sites. *BMC Bioinf.* **2011**, *12* (Suppl 1), S35.

(16) Mattos, C.; Ringe, D. Locating and characterizing binding sites on proteins. *Nat. Biotechnol.* **1996**, *14*, 595−599.

(17) Goodford, P. J. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J. Med. Chem.* **1985**, *28*, 849−857.

(18) Landon, M. R.; Lancia, D. R., Jr.; Yu, J.; Thiel, S. C.; Vajda, S. Identification of hot spots within druggable binding regions by computational solvent mapping of proteins. *J. Med. Chem.* **2007**, *50*, 1231−1240.

(19) Hall, D. R.; Ngan, C. H.; Zerbe, B. S.; Kozakov, D.; Vajda, S. Hot spot analysis for driving the development of hits into leads in fragment-based drug discovery. *J. Chem. Inf. Model.* **2012**, *52*, 199−209.

(20) Friesner, R. A.; Murphy, R. B.; Repasky, M. P.; Frye, L. L.; Greenwood, J. R.; Halgren, T. A.; Sanschagrin, P. C.; Mainz, D. T. Extra precision glide: docking and scoring incorporating a model of

hydrophobic enclosure for protein-ligand complexes. *J. Med. Chem.* **2006**, *49*, 6177−6196.
(21) Smith, B. C.; Denu, J. M. Chemical mechanisms of histone lysine and arginine modifications. *Biochim. Biophys. Acta* **2009**, *1789*, 45−57.
(22) Hajduk, P. J.; Greer, J. A decade of fragment-based drug design: strategic advances and lessons learned. *Nat. Rev. Drug Discov.* **2007**, *6*, 211−219.
(23) Liu, L.; Zhen, S.; Denton, E.; Marsden, B.; Schapira, M. ChromoHub: a data hub for navigators of chromatin-mediated signalling. *Bioinformatics* **2012**, *28*, 2205−2206.
(24) Letunic, I.; Bork, P. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics* **2007**, *23*, 127−128.
(25) Irwin, J. J.; Shoichet, B. K. ZINC–a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* **2005**, *45*, 177−182.
(26) Boobbyer, D. N.; Goodford, P. J.; McWhinnie, P. M.; Wade, R. C. New hydrogen-bond potentials for use in determining energetically favorable binding sites on molecules of known structure. *J. Med. Chem.* **1989**, *32*, 1083−1094.