nature materials

SUPPLEMENTARY INFORMATION

https://doi.org/10.1038/s41563-017-0007-z

In the format provided by the authors and unedited.

Quantitative self-assembly prediction yields targeted nanomedicines

Yosi Shamay[1,2], Janki Shah[1], Mehtap Işık[1,3], Aviram Mizrachi[1,4], Josef Leibold[1], Darjus F. Tschaharganeh[5], Daniel Roxbury[6], Januka Budhathoki-Uprety[1], Karla Nawaly[1], James L. Sugarman[1], Emily Baut[1,7], Michelle R. Neiman[1], Megan Dacek[1,7], Kripa S. Ganesh[1,7], Darren C. Johnson[1,3], Ramya Sridharan[1,7], Eren L. Chu[1,7], Vinagolu K. Rajasekhar[1], Scott W. Lowe[1], John D. Chodera[1] and Daniel A. Heller[1,7]*

[1]Memorial Sloan Kettering Cancer Center, New York, NY, USA. [2]Faculty of Biomedical Engineering, Technion—Israel Institute of Technology, Haifa, Israel. [3]Tri-Institutional PhD Program in Chemical Biology, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [4]Department of Otolaryngology Head and Neck Surgery, Rabin Medical Center and Sackler Faculty of Medicine, Tel Aviv University, Tel Aviv, Israel. [5]Helmholtz-University Group "Cell Plasticity and Epigenetic Remodeling", German Cancer Research Center (DKFZ) & Institute of Pathology University Hospital, Heidelberg, Germany. [6]Department of Chemical Engineering, University of Rhode Island, Kingston, RI 02881, USA. [7]Weill Cornell Medical College, Cornell University, New York, NY, USA.
*e-mail: hellerd@mskcc.org

Quantitative Self-Assembly Prediction Yields Targeted Nanomedicines

Yosi Shamay[1,2], Janki Shah[1], Mehtap Işık[1,3], Aviram Mizrachi[1,4], Josef Leibold[1], Darjus F. Tschaharganeh[5], Daniel Roxbury[6], Januka Budhathoki-Uprety[1], Karla Nawaly[1], James L. Sugarman[1], Emily Baut[1,7], Michelle R. Neiman[1], Megan Dacek[1,7], Kripa S. Ganesh[1,7], Darren C. Johnson[1,3], Ramya Sridharan[1,7], **E**ren L. Chu[1,7], Vinagolu K. Rajasekhar[1], Scott W. Lowe[1], John D. Chodera[1], Daniel A. Heller[1,7*]

[1]Memorial Sloan Kettering Cancer Center, New York, NY

[2]Faculty of Biomedical Engineering, Technion - Israel Institute of Technology, Haifa, Israel

[3]Tri-Institutional PhD Program in Chemical Biology, Memorial Sloan Kettering Cancer Center, New York, NY

[4]Department of Otolaryngology Head and Neck Surgery, Rabin Medical Center and Sackler Faculty of Medicine, Tel Aviv University, Tel Aviv, Israel

[5]Helmholtz-University Group "Cell Plasticity and Epigenetic Remodeling", German Cancer Research Center (DKFZ) & Institute of Pathology University Hospital, Heidelberg, Germany

[6]Department of Chemical Engineering, University of Rhode Island, Kingston, RI 02881

[7]Weill Cornell Medical College, Cornell University, New York, NY

Supplementary Discussion

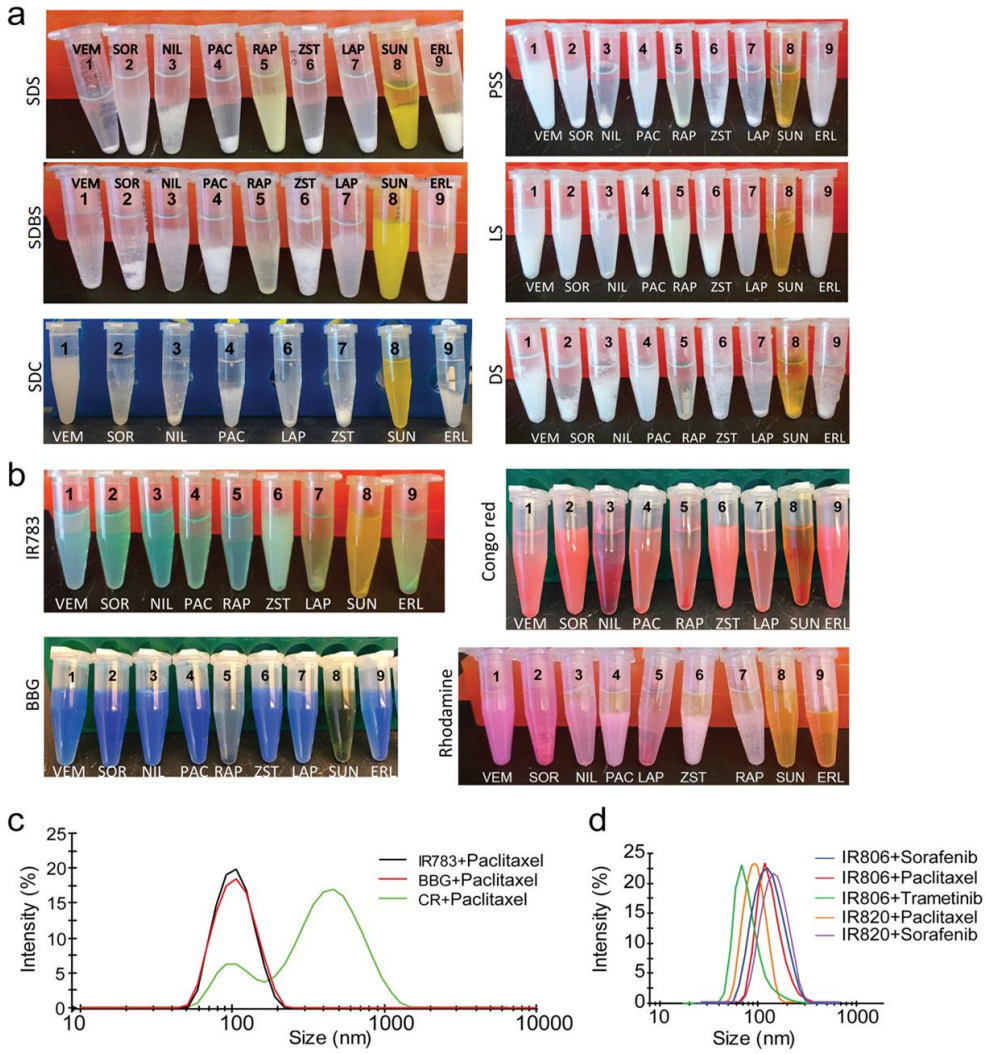
Molecular Dynamics Modeling of Self Assembly

Clustering of the REMD trajectory was used to determine the most populous conformation in the simulation. Accounting for an initial equilibration period, the final 25 ns of the 300 K replica trajectory (temperature at which the experiment was performed) was used for all analysis. A native Gromacs clustering algorithm (g_cluster) was used with a root mean square deviation (RMSD) cutoff of 1.2 nm based upon the spatial positions of the drug atoms. The top cluster from the 5,000 available snapshots represented 9.6% and 0.8% of the trajectory for the Sorafenib and Taselisib simulations, respectively (Main text, **Figure 2h**). The significantly lower percentage of trajectory in the top cluster in the Taselisib simulation suggests an intrinsically more random preferred conformation for this drug-dye combination. Normalized radial particle density histograms were constructed from the top cluster configurations (**Supplementary Figure 12c**).

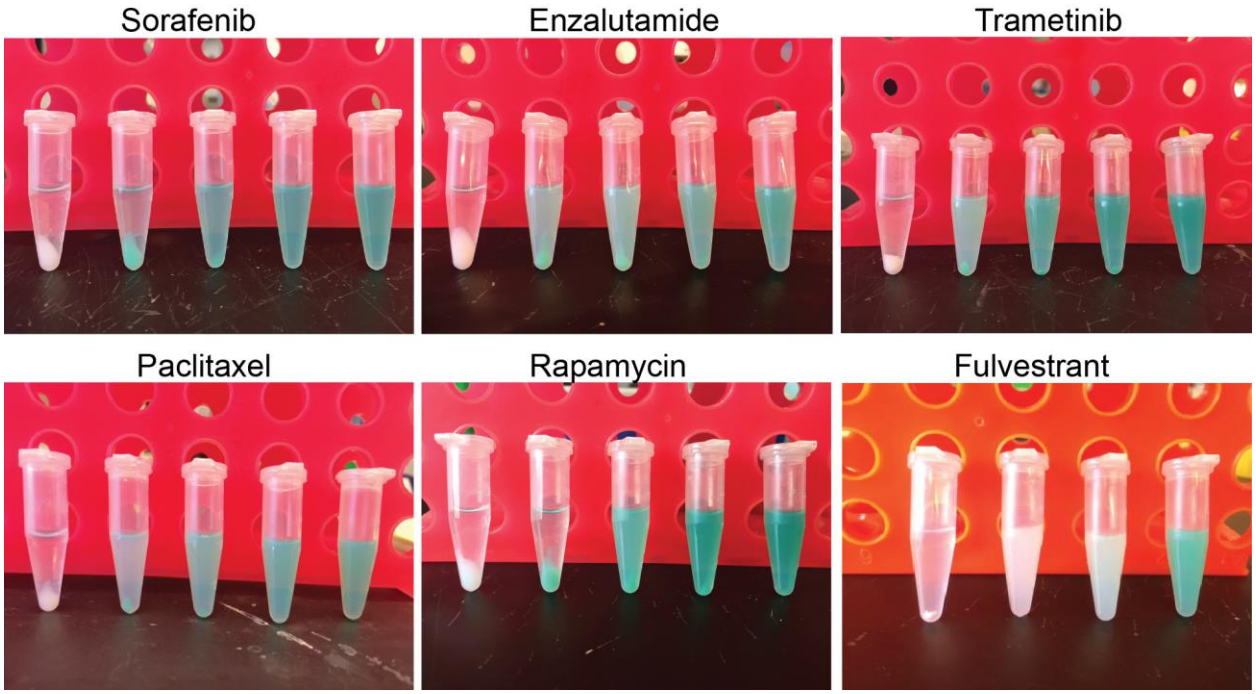
The solvent accessibilities to the surfaces of the drugs were analyzed to determine accessibility in the complexes. Water and ion accessibilities were analyzed using the Gromacs function 'g_sas'. In order to compare across the two simulations with differing drug surface areas, we quantified the amount of exposed drug to the solvent with the dye present, and additionally with the dye removed from the trajectory. The percentage change in solvent accessible drug surface area was quantified, revealing that the dye shields the Sorafenib significantly more than Taselisib, 27.9 ± 3.1% vs. 20.3 ± 3.7% (**Supplementary Figure 12d**).

Hydrogen bonding analysis was performed using the Gromacs function 'g_hbond'. We calculated the total number of hydrogen bonds between solute molecules in the system, and between dye and drug molecules (**Supplementary Figure 12e-f**). The dye was not able to hydrogen bond to itself, and thus the total number of bonds comprised drug-drug and dye-drug interactions. The average number of total hydrogen bonds was 13.3 ± 2.7 and 1.9 ± 1.4 for Sorafenib and Taselisib simulations, respectively. Moreover, the number of dye-drug hydrogen bonds was 10.3 ± 2.6 and 0.4 ± 0.6 for IR783-Sorafenib and IR783-Taselisib, respectively.

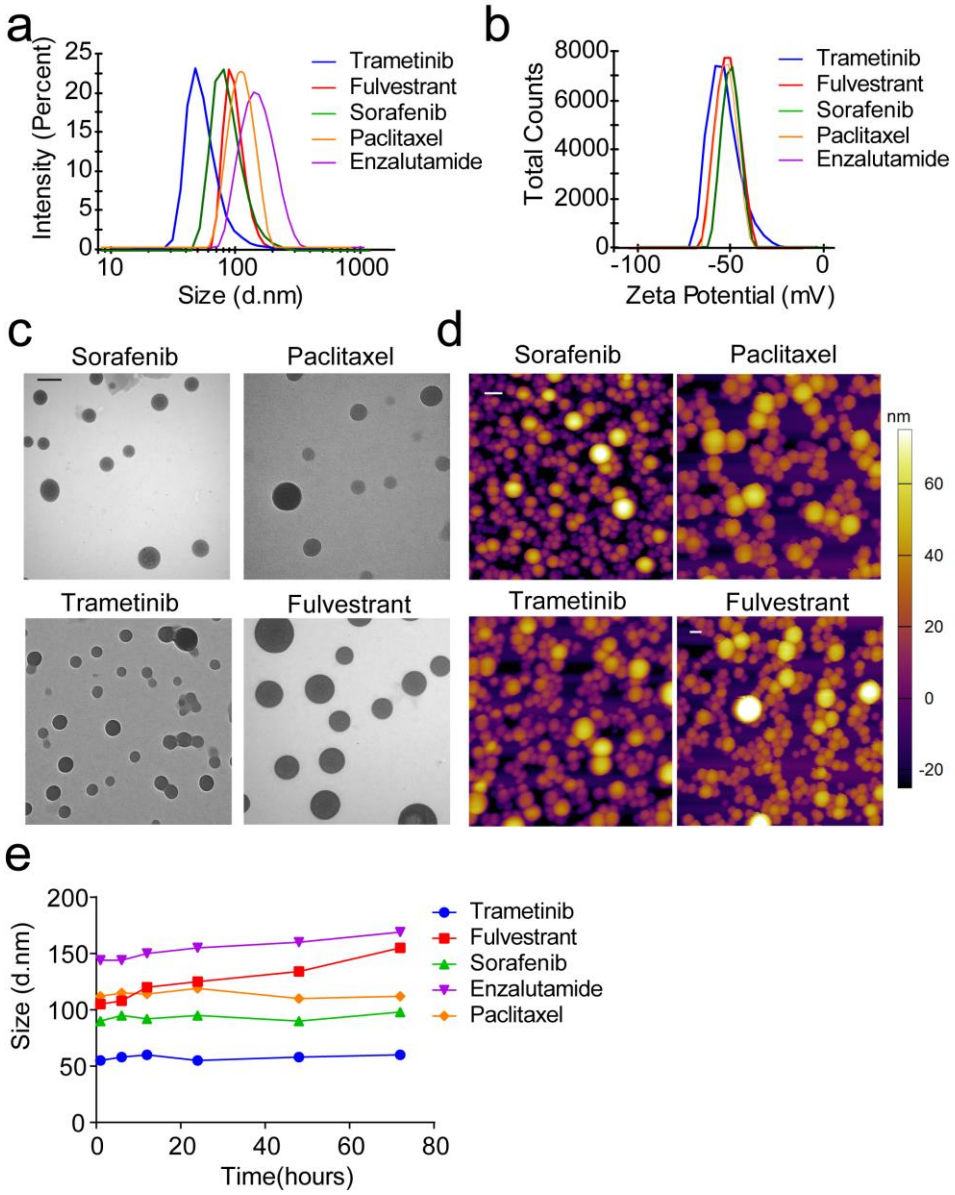
Supplementary Figures



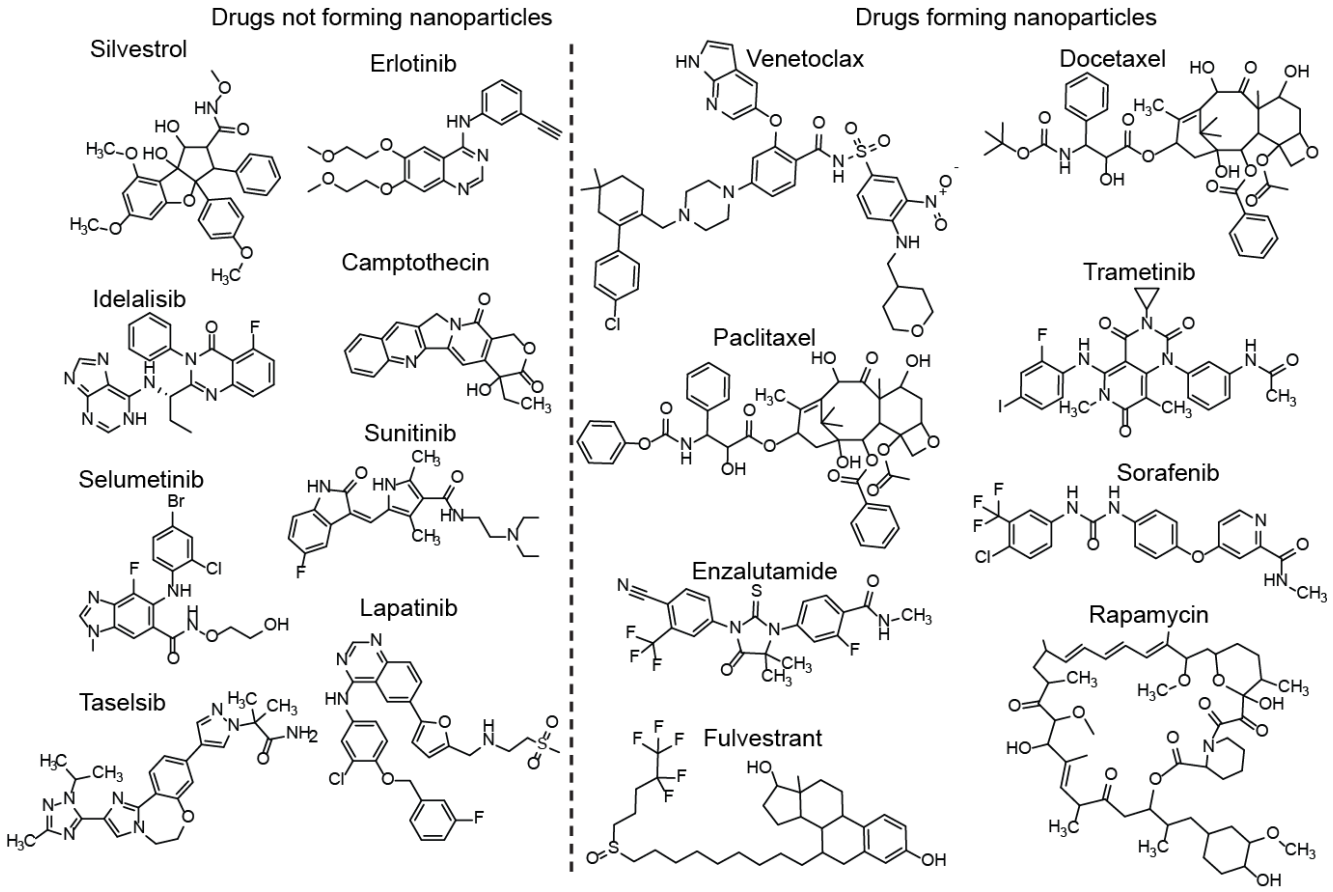
Supplementary Figure 1. Excipient screen for drug suspension. (a) Excipient mixtures with nine different drugs: 1) VEM=Vemurafinib, 2) SOR=Sorafenib, 3) NIL=Nilotinib, 4) PAC=Paclitaxel, 5) RAP=Rapamycin, 6) ZST=ZSTK474, 7) LAP=Lapatinib, 8) SUN=Sunitinib, 9) ERL=Erlotinib. Anionic amphiphilic compounds: sodium dodecyl sulfate (SDS), sodium dodecylbenzene sulfonate (SDBS), sodium deoxycholate (SDC), and three anionic poly electrolytes: poly-4-styrensulfonate (PSS), lignin sulfonate (LS) and dextran sulfate (DS). **(b)** Mixtures of the nine drugs with dye excipients IR783, Brilliant Blue G (BBG), rhodamine, and Congo red. **(c)** DLS measurements of the paclitaxel suspensions in IR783, Brilliant Blue G, or Congo Red. **(d)** DLS measurements of drug suspensions using other indocyanine dyes IR806 and IR820.



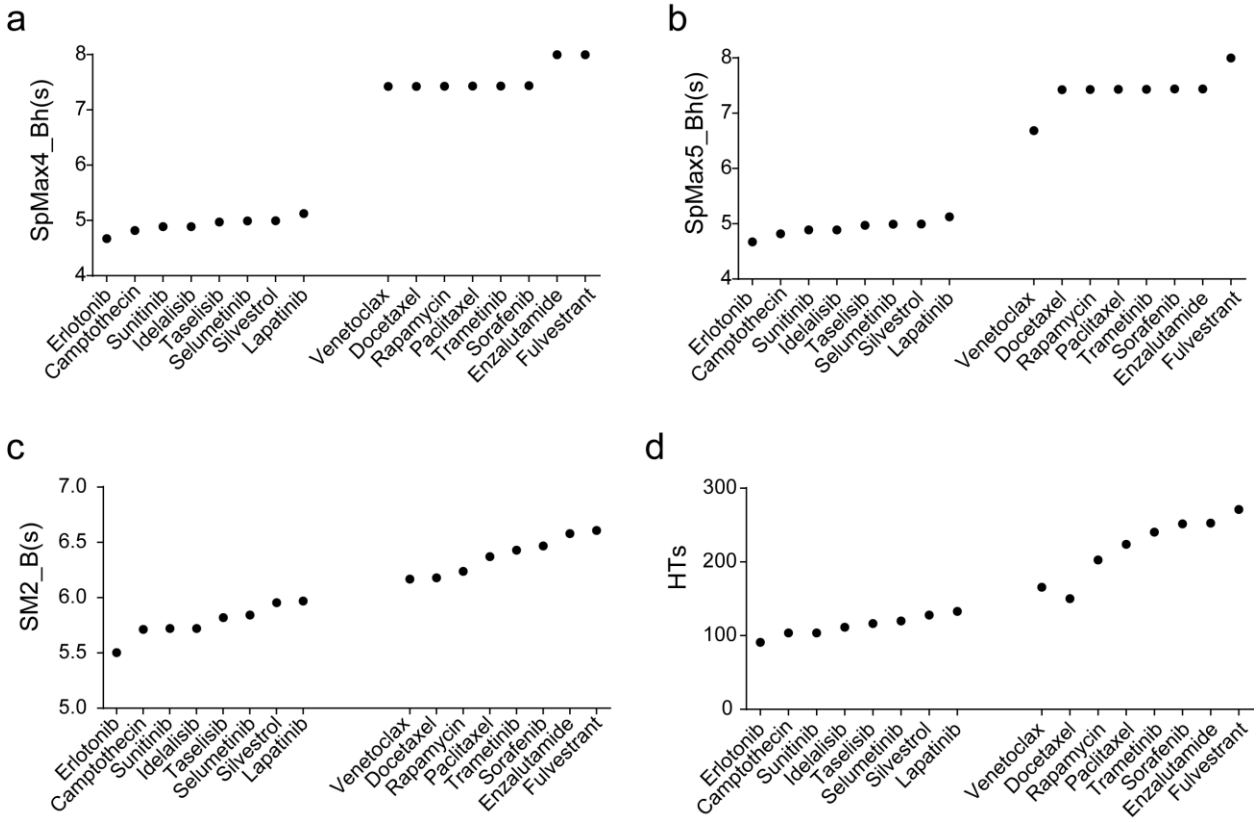
Supplementary Figure 2. **Indocyanine-drug suspensions.** Drug/indocyanine mixtures at increasing drug:indocyanine ratios. The drug concentration was fixed at 2 mg/ml and the indocyanine IR783 concentrations were 0, 0.005, 0.01, 0.03 and 0.1 mg/ml (from left to right). In the case of fulvestrant, indocyanine concentrations were 0, 0.0001, 0.005 and 0.001 mg/ml. All tubes were centrifuged for 1 min at 3000 g before imaging.



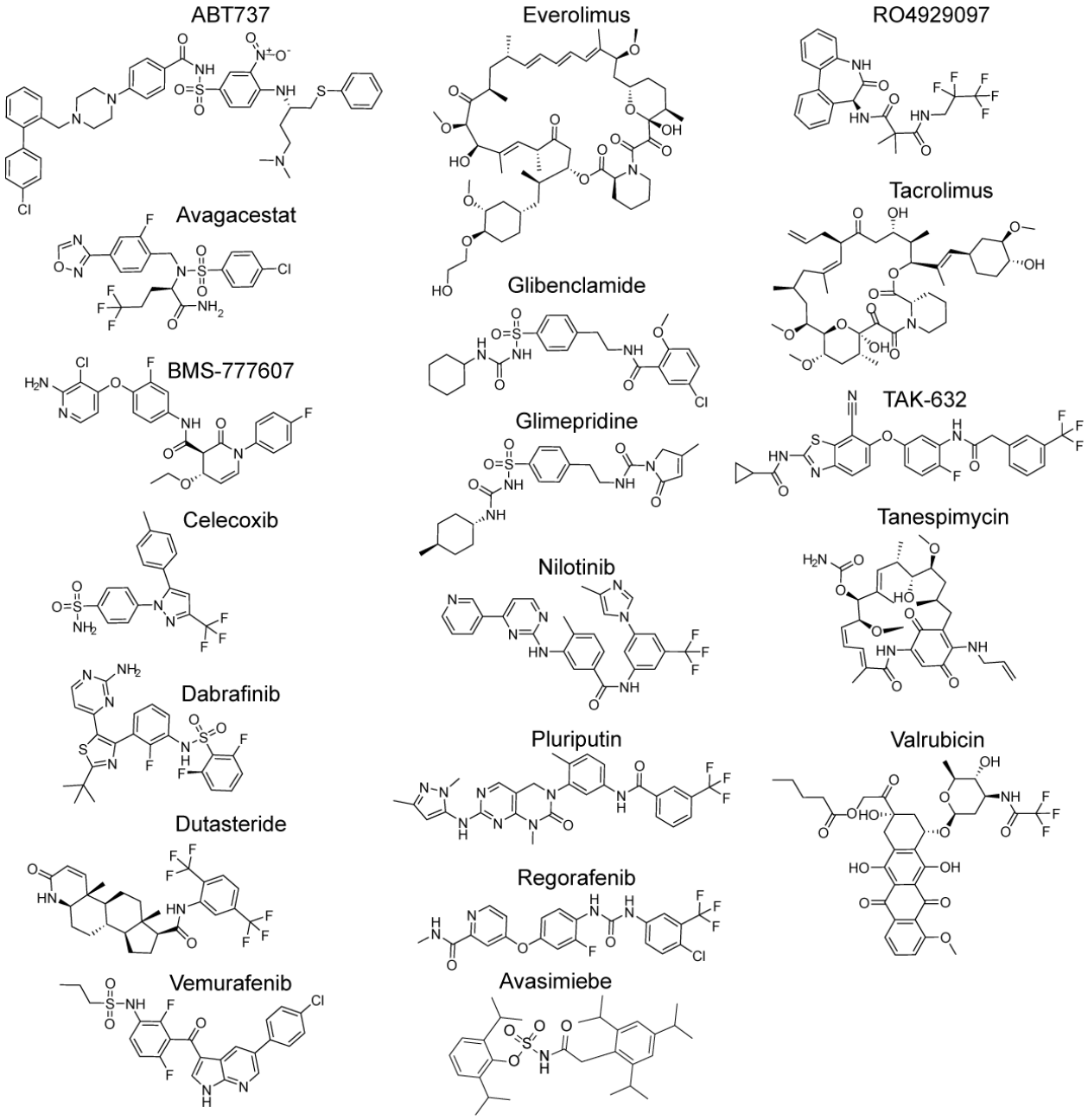
Supplementary Figure 3. **Characterization of IR783 indocyanine nanoparticles. (a)** Nanoparticle diameters measured with dynamic light scattering (DLS). **(b)** Nanoparticle zeta potential measured with electrophoretic light scattering. **(c)** Transmission electron microscopy (TEM) images of indocyanine nanoparticles. Scale Bar = 100 nm. **(d)** Atomic force microscopy (AFM) images of indocyanine nanoparticles. Scale Bars = 100 nm. **(e)** Nanoparticle stability in growth medium containing serum, evaluated by DLS.



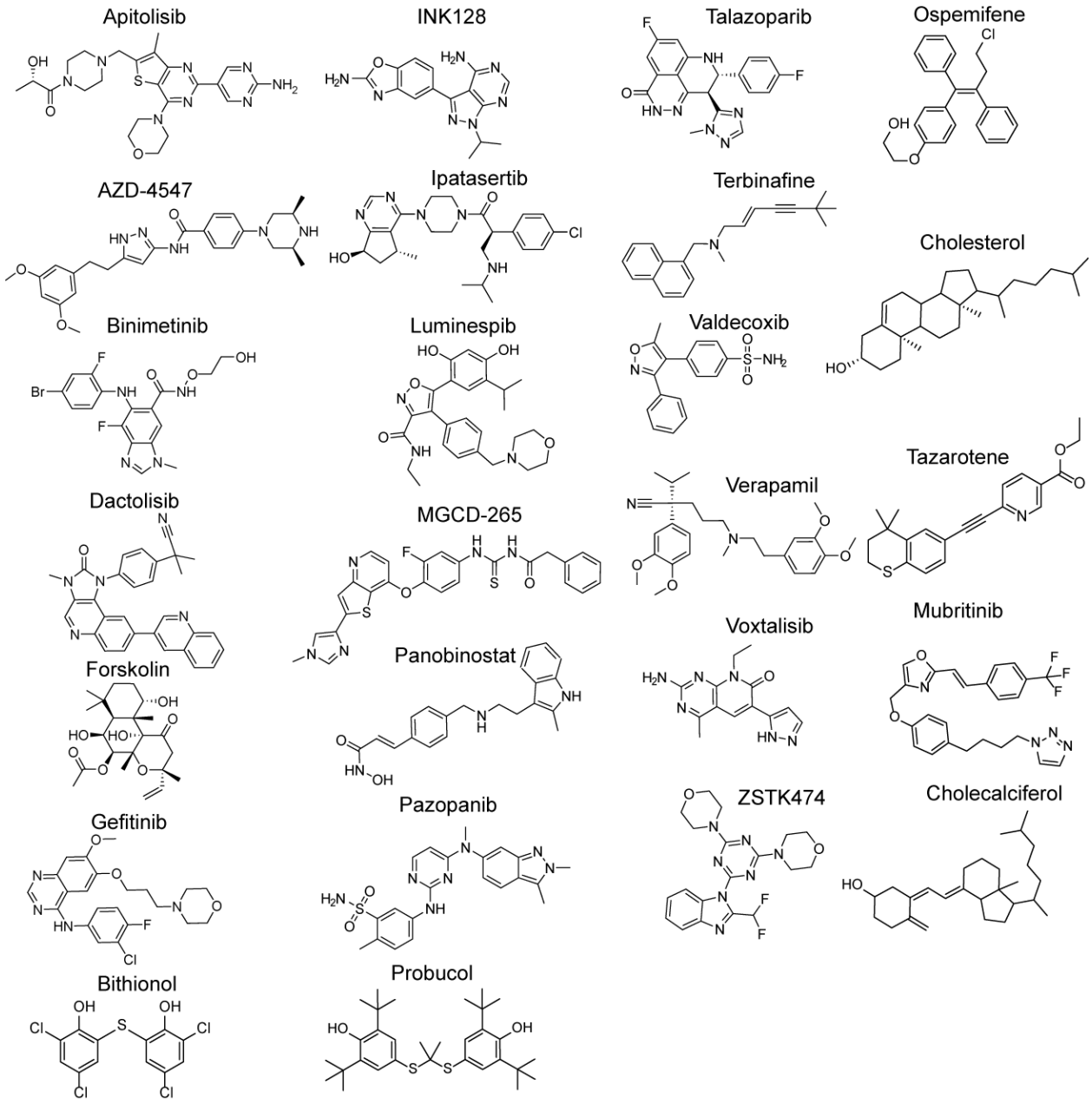
Supplementary Figure 4. **QSNAP model 1 training set.** Molecular structures of drugs used in the training set for INP formation.



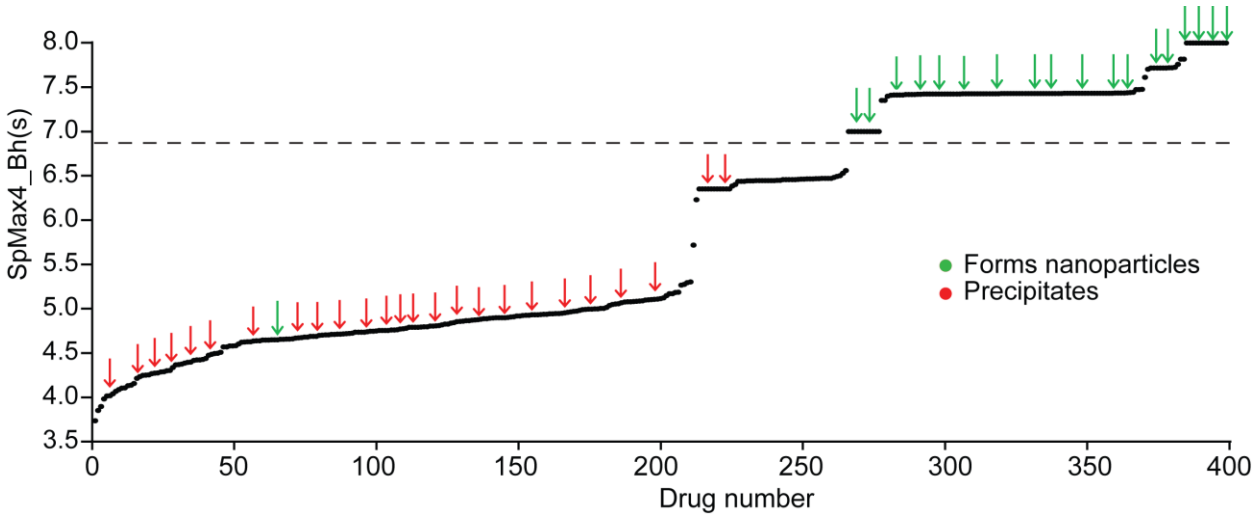
Supplementary Figure 5. **Molecular descriptors of drug suspension.** Four molecular descriptors exhibiting high correlations ($r > 0.85$) with experimental data are shown. In each panel, the first grouping of drugs, on the left, was found experimentally to precipitate, while the second grouping of drugs, on the right, was found to suspend with IR783. **a)** SpMAX4_Bh(s) = largest eigenvalue n. 4 of Burden matrix weighted by I-state. **b)** SpMAX5_Bh(s) = largest eigenvalue n. 5 of Burden matrix weighted by I-state. **c)** SM2_B(s) = spectral moment of order 2 from Burden matrix weighted by I-State. **d)** HTs = H total index weighted by I-state. Calculations were performed using the Dragon 6 software (*Talette*).



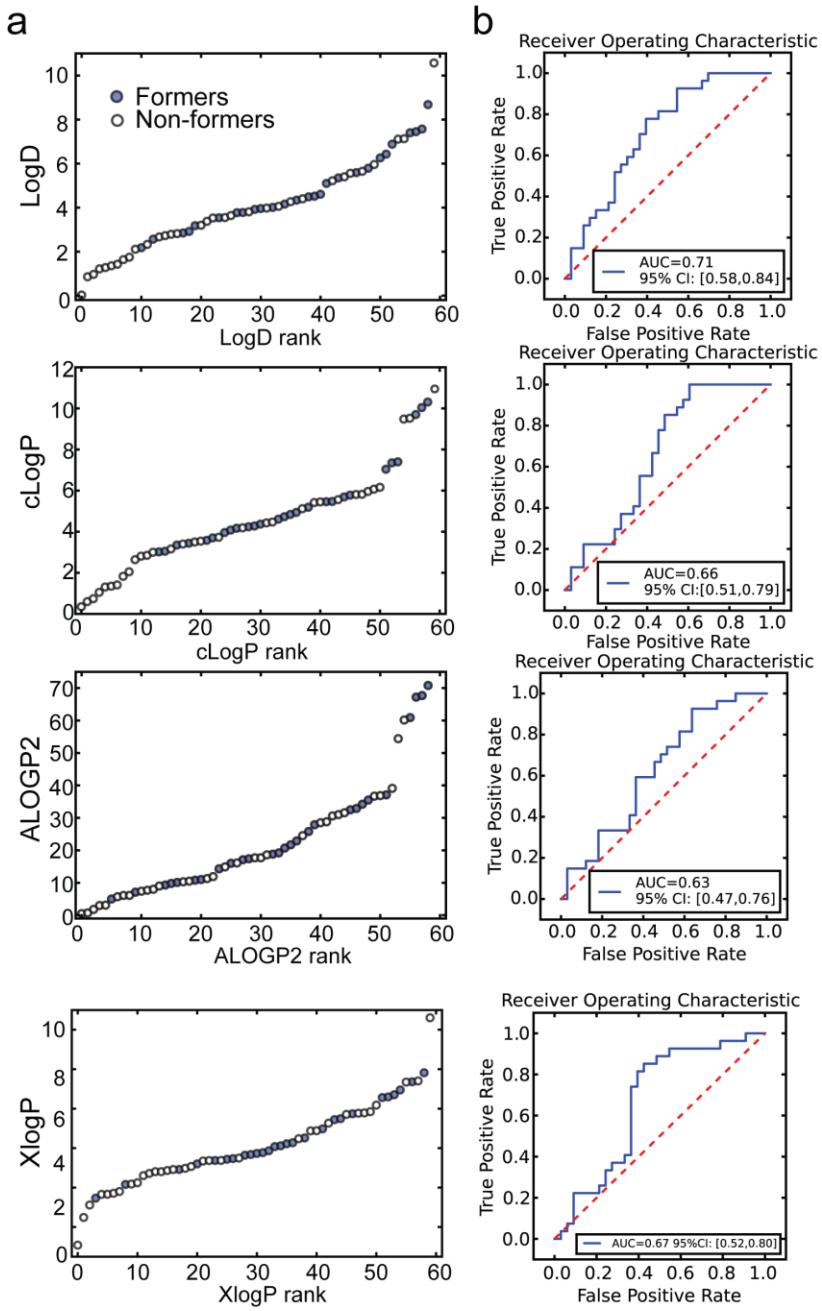
Supplementary Figure 6. Molecular structures of INP-forming drugs selected for prospective validation set.



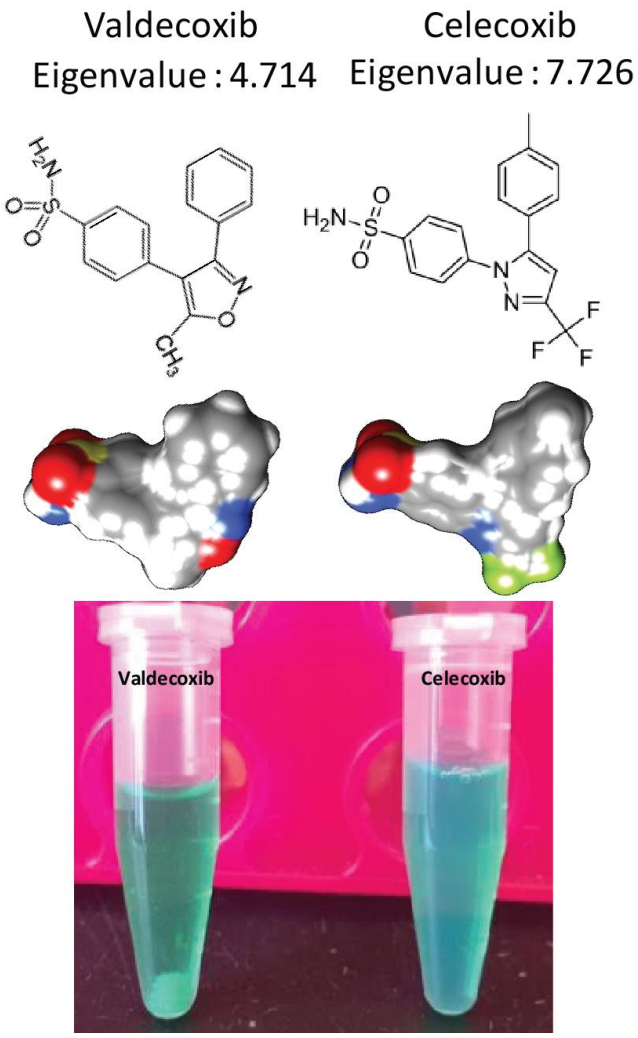
Supplementary Figure 7. Molecular structures of non-INP-forming drugs selected for prospective validation set.



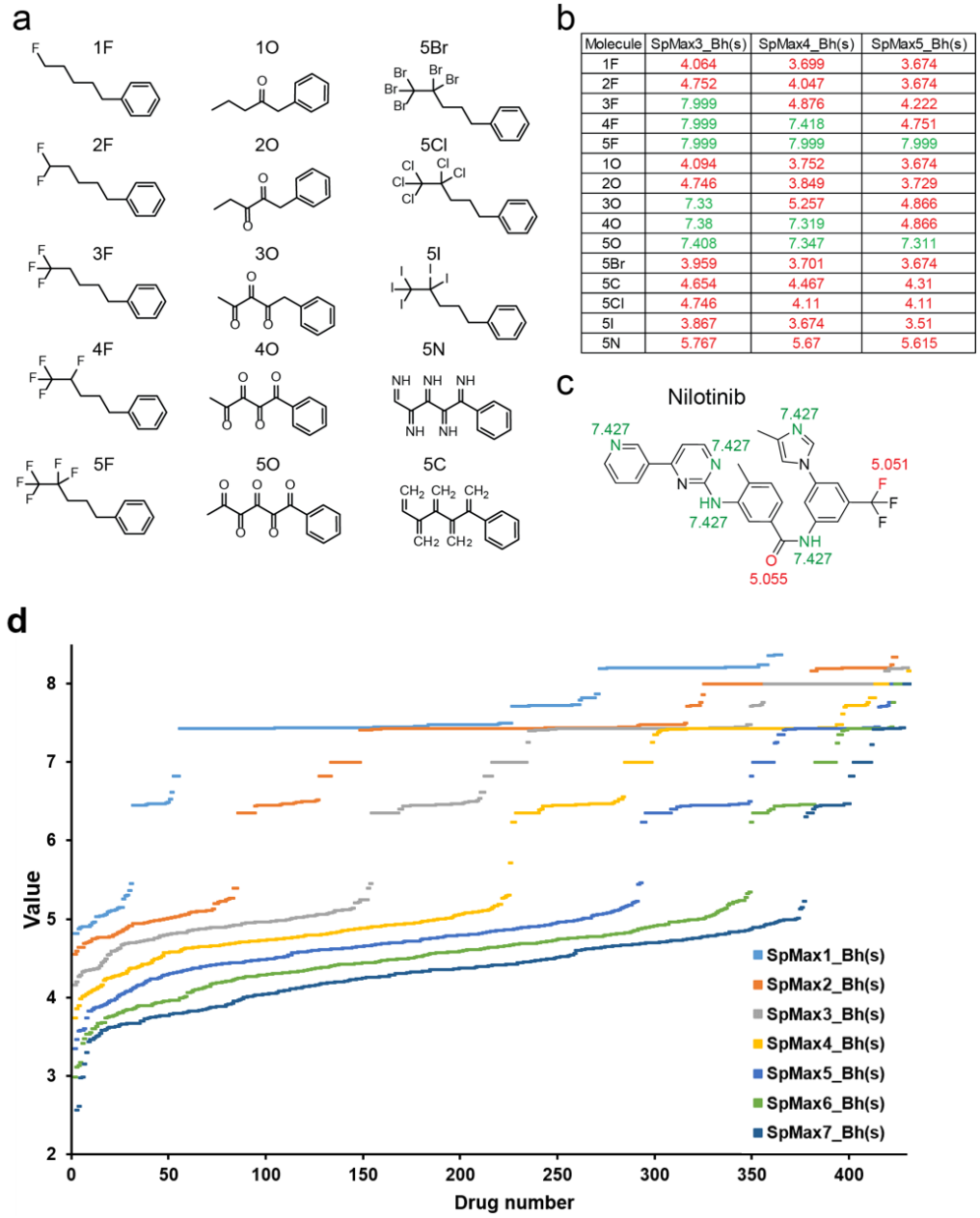
Supplementary Figure 8. SpMAX4_Bh(s) eigenvalues of 400 drugs and experimental validation of 44 drugs. Arrows indicate experimental validation of drug suspension using IR783. Red = drugs that precipitated with indocyanine. Green=drugs that formed stable nanoparticles with indocyanine.



Supplementary Figure 9. Statistical analyses of descriptors of hydrophobicity for all experimentally validated drugs. Descriptor scores plotted in ascending order of the relative descriptor score of each drug (left), and receiver operating characteristic (ROC) curves for the prediction of IR783-mediated drug suspension by each descriptor (right). Hydrophobicity descriptors are LogD at pH 7.4 (Chemicalize), cLogP (Chem 3D), ALOGP2 (Dragon 6) and XLogP (OpenEye).

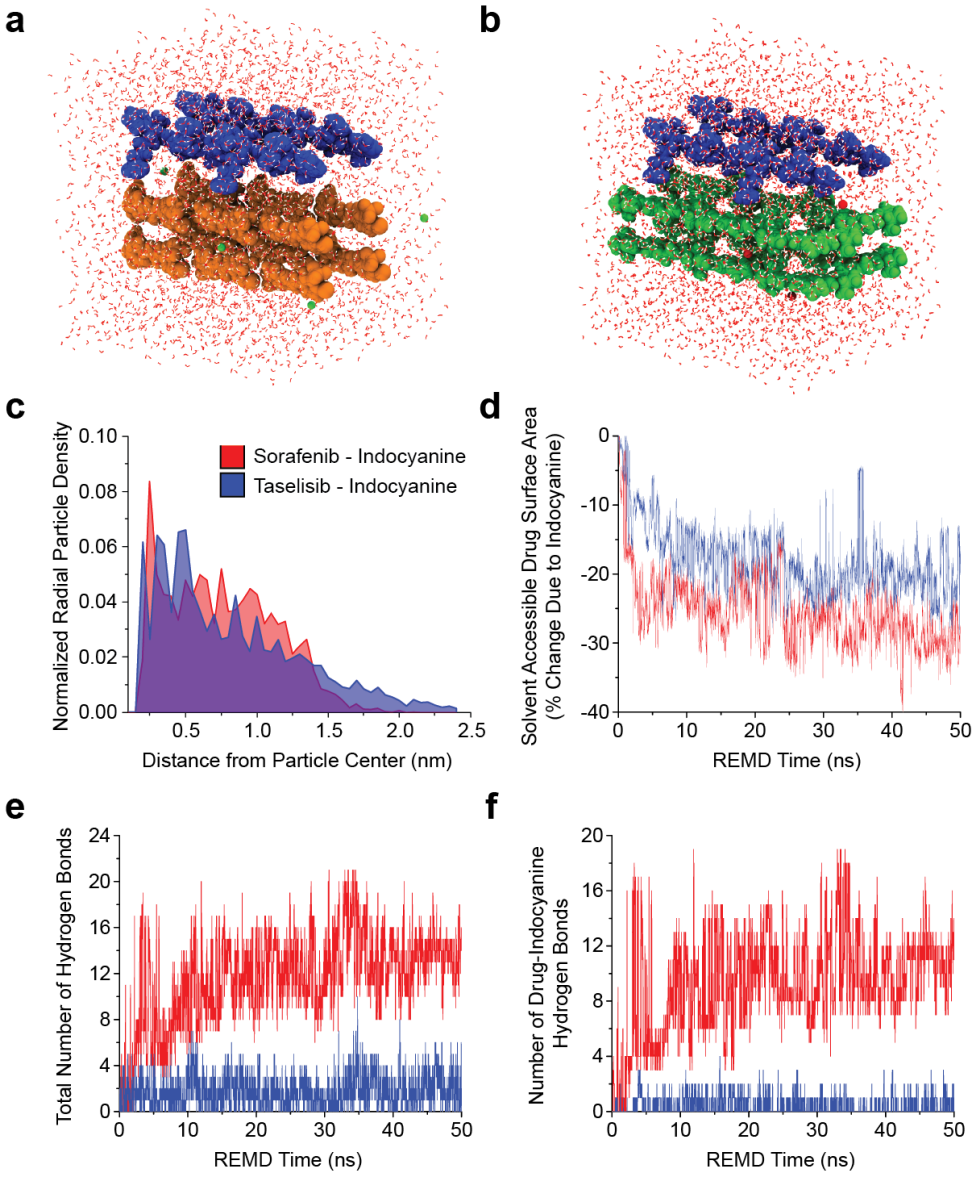


Supplementary Figure 10. **Contribution of fluorine to drug suspension via indocyanine**. Structures of celecoxib and valdexoxib with their SpMAX4_Bh(s) values, and photographs of drug-indocyanine mixtures.

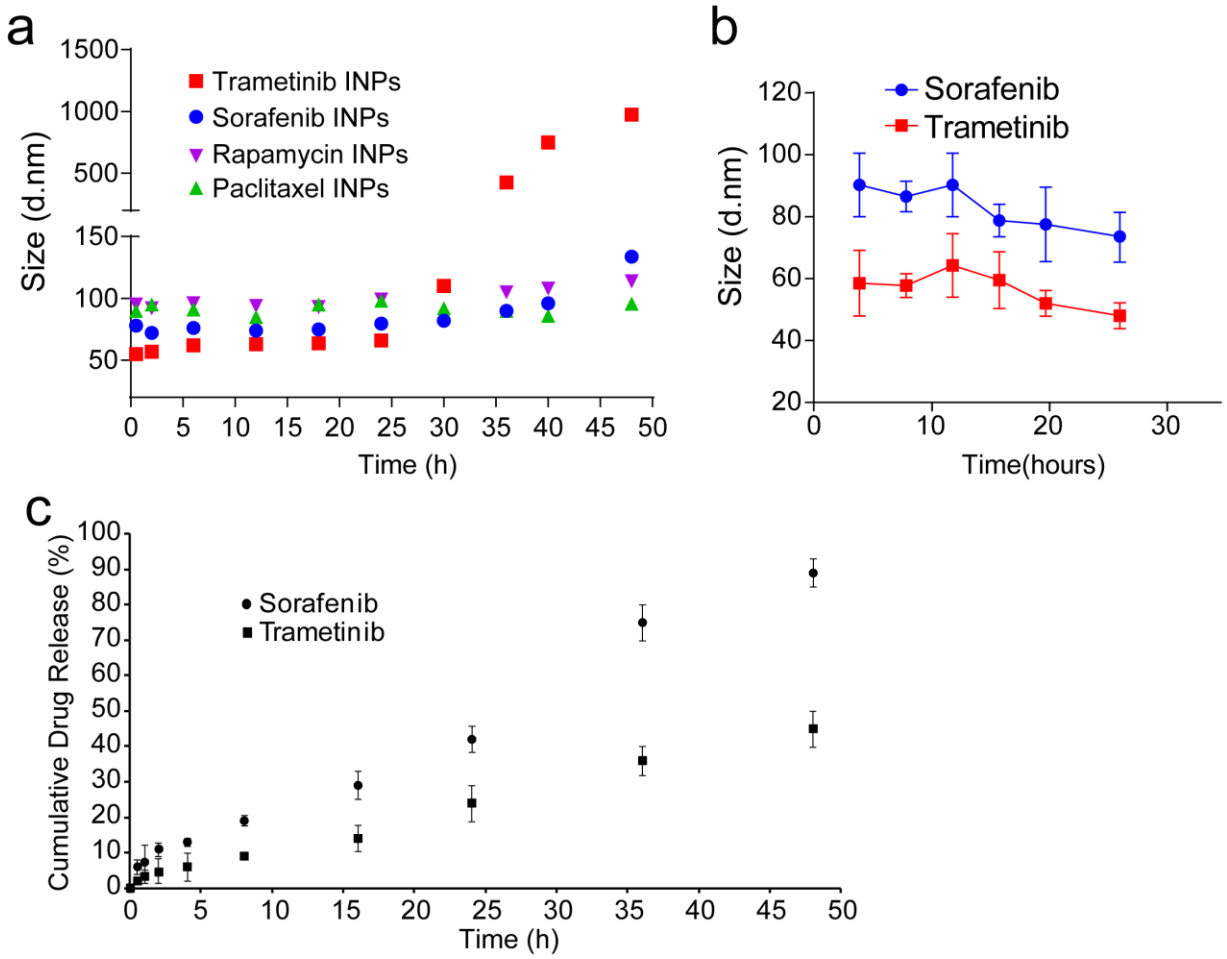


Molecule	SpMax3_Bh(s)	SpMax4_Bh(s)	SpMax5_Bh(s)
1F	4.064	3.699	3.674
2F	4.752	4.047	3.674
3F	7.999	4.876	4.222
4F	7.999	7.418	4.751
5F	7.999	7.999	7.999
1O	4.094	3.752	3.674
2O	4.746	3.849	3.729
3O	7.33	5.257	4.866
4O	7.38	7.319	4.866
5O	7.408	7.347	7.311
5Br	3.959	3.701	3.674
5C	4.654	4.467	4.31
5Cl	4.746	4.11	4.11
5I	3.867	3.674	3.51
5N	5.767	5.67	5.615

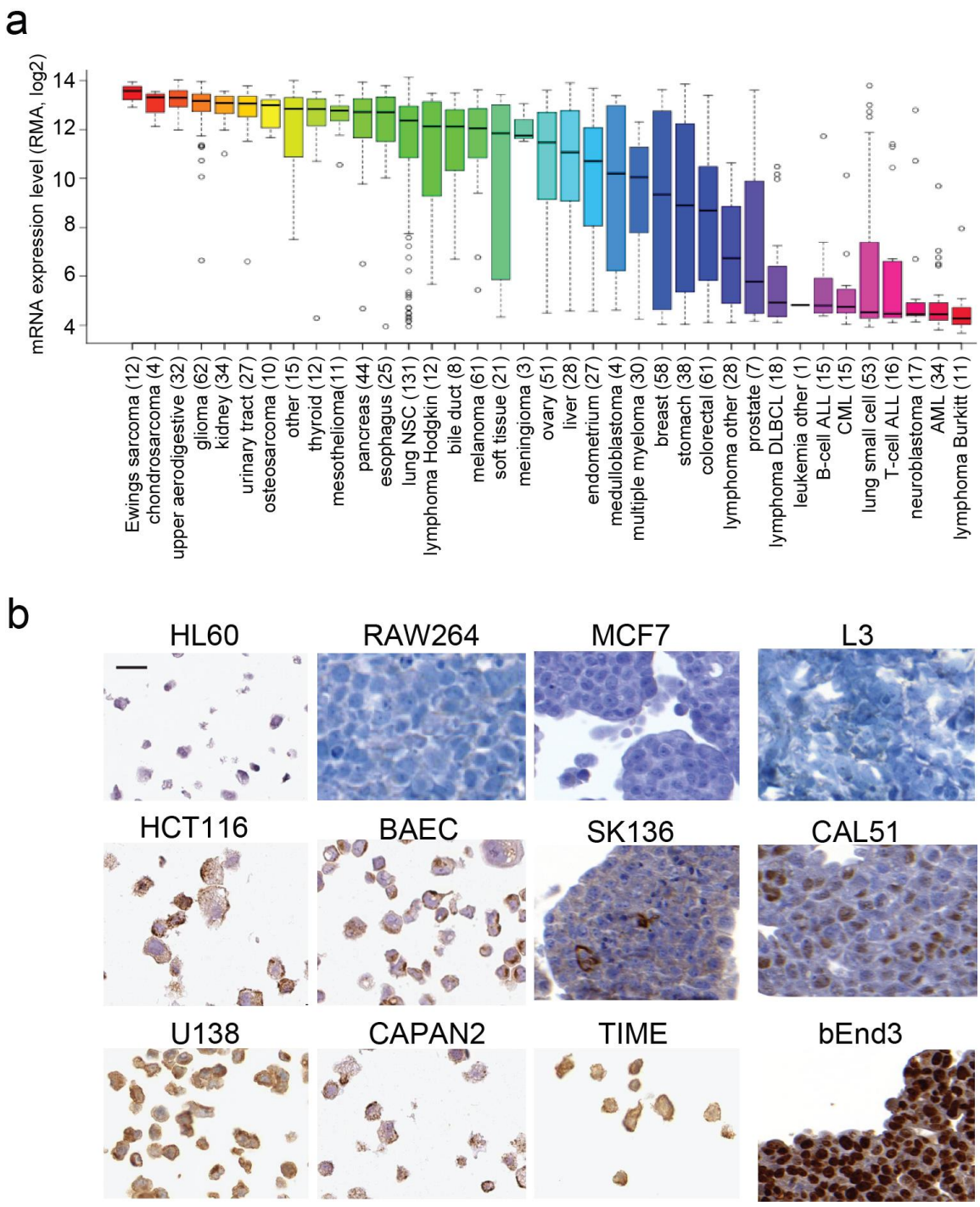
Supplementary Figure 11. Investigation of Burden matrix eigenvalues. a) Systematic addition/deletion of functional groups of molecular structures. **b)** Calculated Burden matrix eigenvalues SpMAX3_Bh(s), SpMAX4_Bh(s), and SpMAX5_Bh(s) corresponding to the molecular structures in (a). Red = values below 7. Green = values above 7. **c)** SpMAX4_Bh(s) values of the drug nilotinib after single heteroatom (colored red or green) replacements with carbon atoms. Red = SpMAX4_Bh(s) values below 7. Green = values above 7. **d**) Seven Burden matrix eigenvalues for 430 molecules, sorted by value. Calculations were performed with Dragon 6 software.



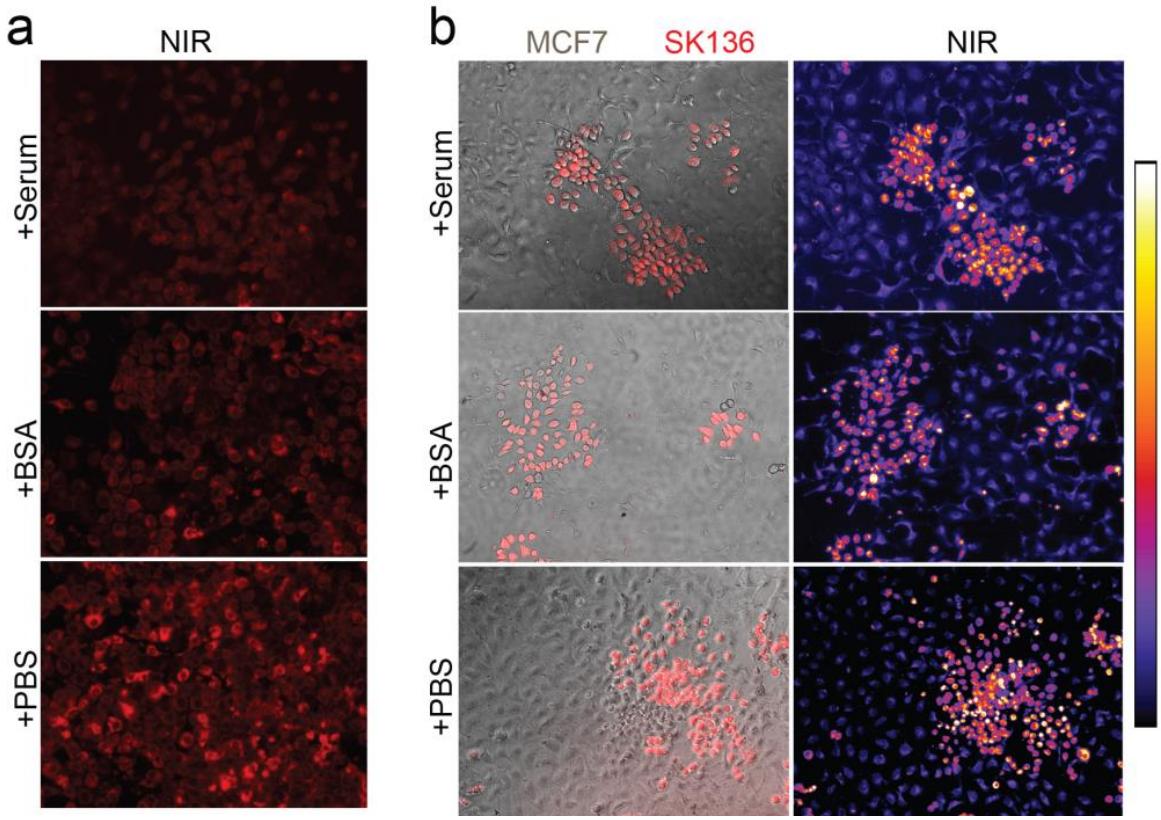
Supplementary Figure 12. **Molecular dynamics simulations of indocyanine nanoparticles.** Initial configurations for **(a)** sorafenib-indocyanine and **(b)** taselisib-indocyanine simulations, each containing 12 drug molecules, 4 dye molecules, counterions, and water. Indocyanine molecules are blue, sorafenib is orange, and taselisib is green. **(c)** Normalized radial particle density histograms for the most probable configurations of sorafenib-indocyanine and taselisib-indocyanine, plotted as a function of distance from the particle center. **(d)** Change in solvent accessible drug surface area plotted as a function of REMD time. **(e)** Total number of intra-nanoparticle (non-water) hydrogen bonds plotted as a function of REMD time. **(f)** Number of drug-indocyanine hydrogen bonds plotted as a function of REMD time. In (c) through (f), sorafenib-indocyanine data is denoted by red curves and taselisib-indocyanine data is blue.



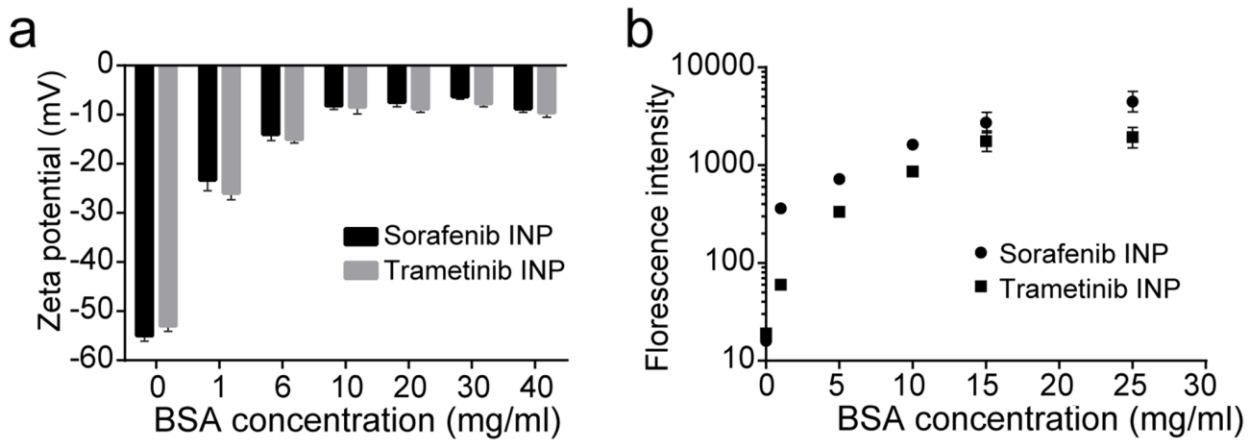
Supplementary Figure 13. **Nanoparticle stability and drug release. (a)** Stability of INPs encapsulating four different drugs at high dilution (sink conditions). 1 mg/ml of nanoparticle solution was diluted 10000x in PBS (pH = 7.4). Nanoparticle diameter was measured by DLS. **(b)** Nanoparticle stability in 100% mouse plasma, evaluated by DLS. **(c)** Drug release profiles of sorafenib and trametinib from their respective nanoparticles in PBS. Error bars are standard deviation of the mean.



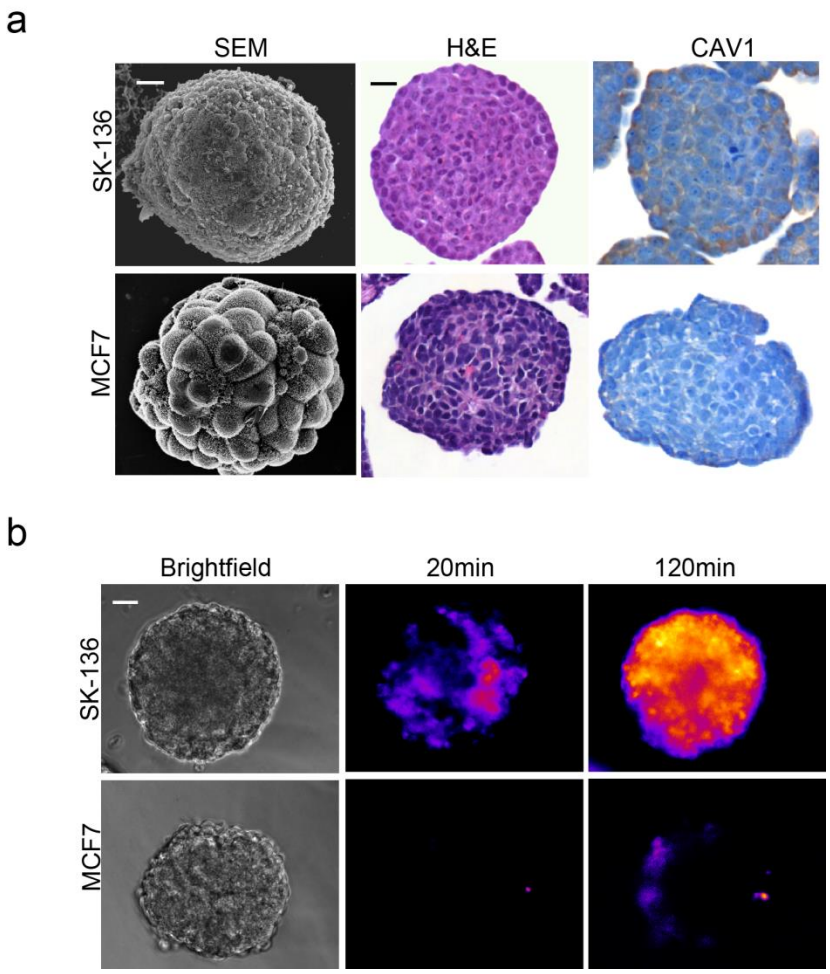
Supplementary Figure 14. Differential expression of CAV1 in cell lines. (**a**) CAV1 gene expression profile in human cell lines, obtained from CCLE database (Entrez ID: 857). (**b**) Immunohistochemical staining for CAV1 in cell lines. Scale bar = 50 µM. Images of HL60, U138, and TIME cell lines were obtained from the Human Protein Atlas database.



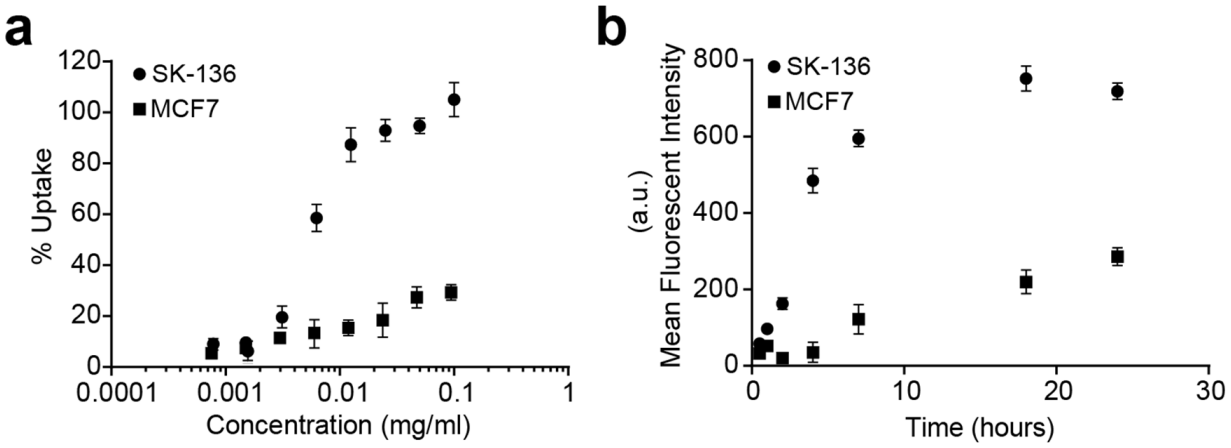
Supplementary Figure 15. Serum and albumin effects on nanoparticle uptake. (a) Fluorescence microscopy of SK-136 cells administered with sorafenib INPs (incubated at 50 µg/ml for 1h) in DMEM containing either 10% serum, 45 mg/ml BSA, or PBS as control. **(b)** Differential uptake of sorafenib INP (50 µg/ml for 1h) in co-culture of SK-136 and SKOV3 in DMEM media containing 10% serum, 45 mg/ml BSA, or PBS as control.



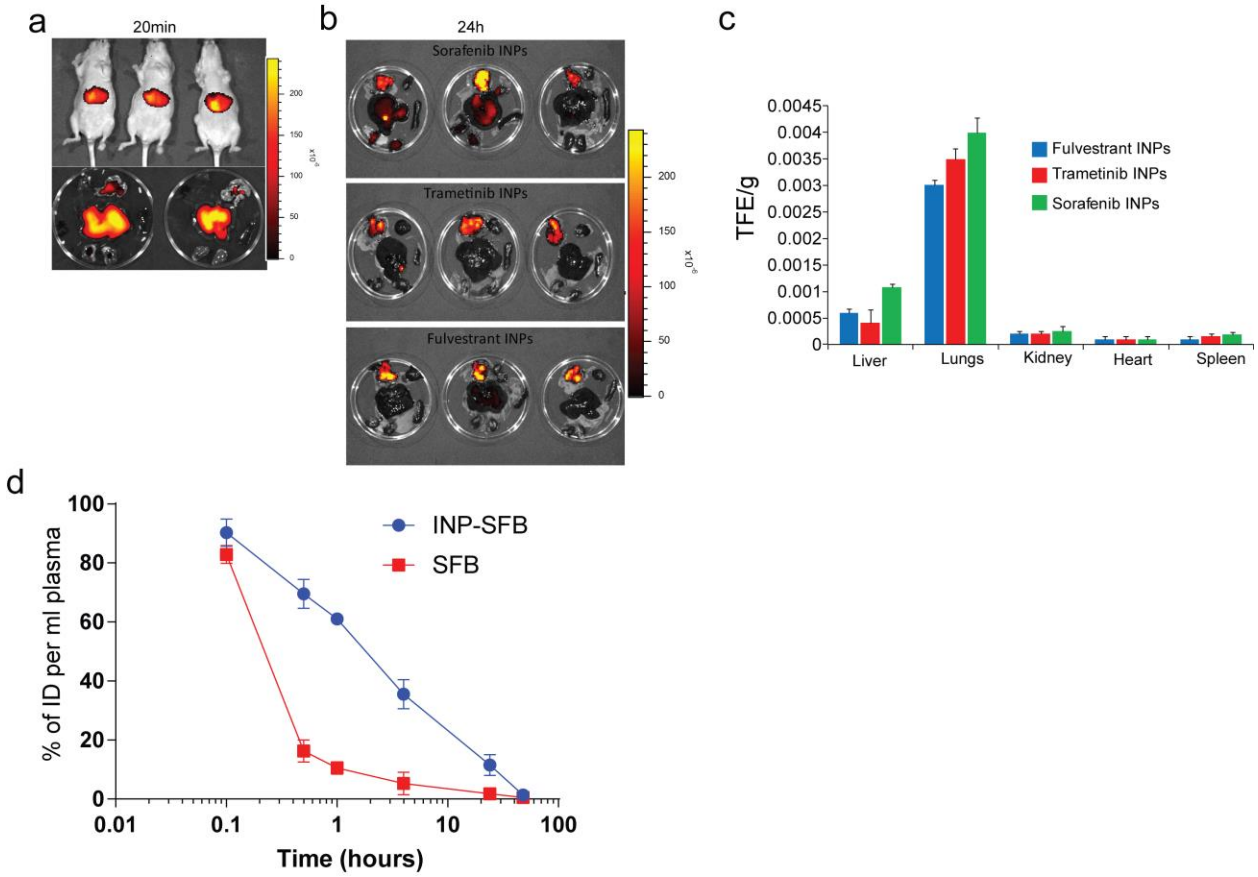
Supplementary Figure 16. Protein adsorption measurements. (a) Electrophoretic light scattering zeta potential measurements of sorafenib INPs and trametinib INPs in increasing concentrations of BSA. **(b)** Fluorescence intensity of BSA-FITC upon incubation at increasing concentrations with sorafenib INPs or trametinib INPs. Nanoparticles (0.5 mg/ml) were incubated with BSA-FITC for 15 min and then centrifuged at 20,000 rcf for 5 min. Samples were suspended in PBS before measuring fluorescence intensity. Error bars are standard deviation of the mean.



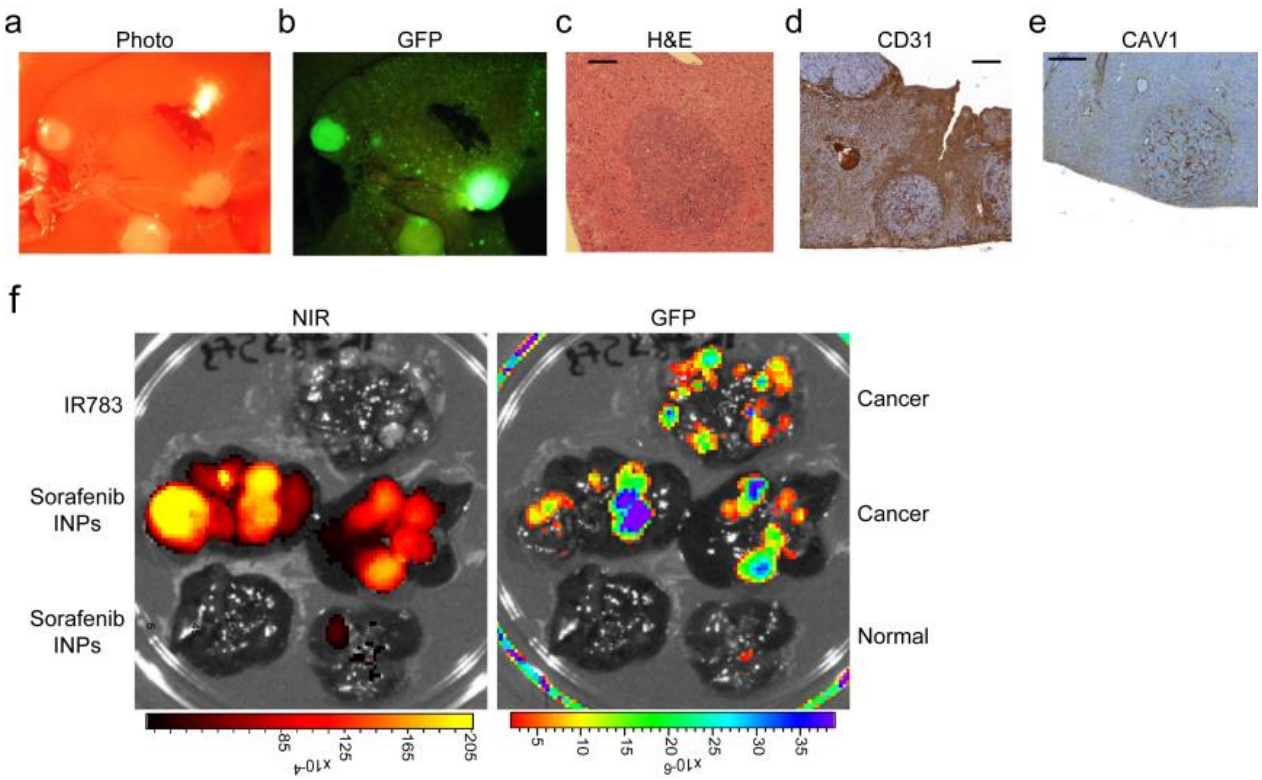
Supplementary Figure 17. **Indocyanine nanoparticle targeting of tumor spheroids. (a)** Characterization of tumor spheres with SEM, H&E stain, and immunohistochemical stain for CAV1. **(b)** Fluorescence microscopy of near-infrared dye emission of tumor spheroids after 20 min and 120 min of incubation with sorafenib INPs. Scale bar = 20 μm.



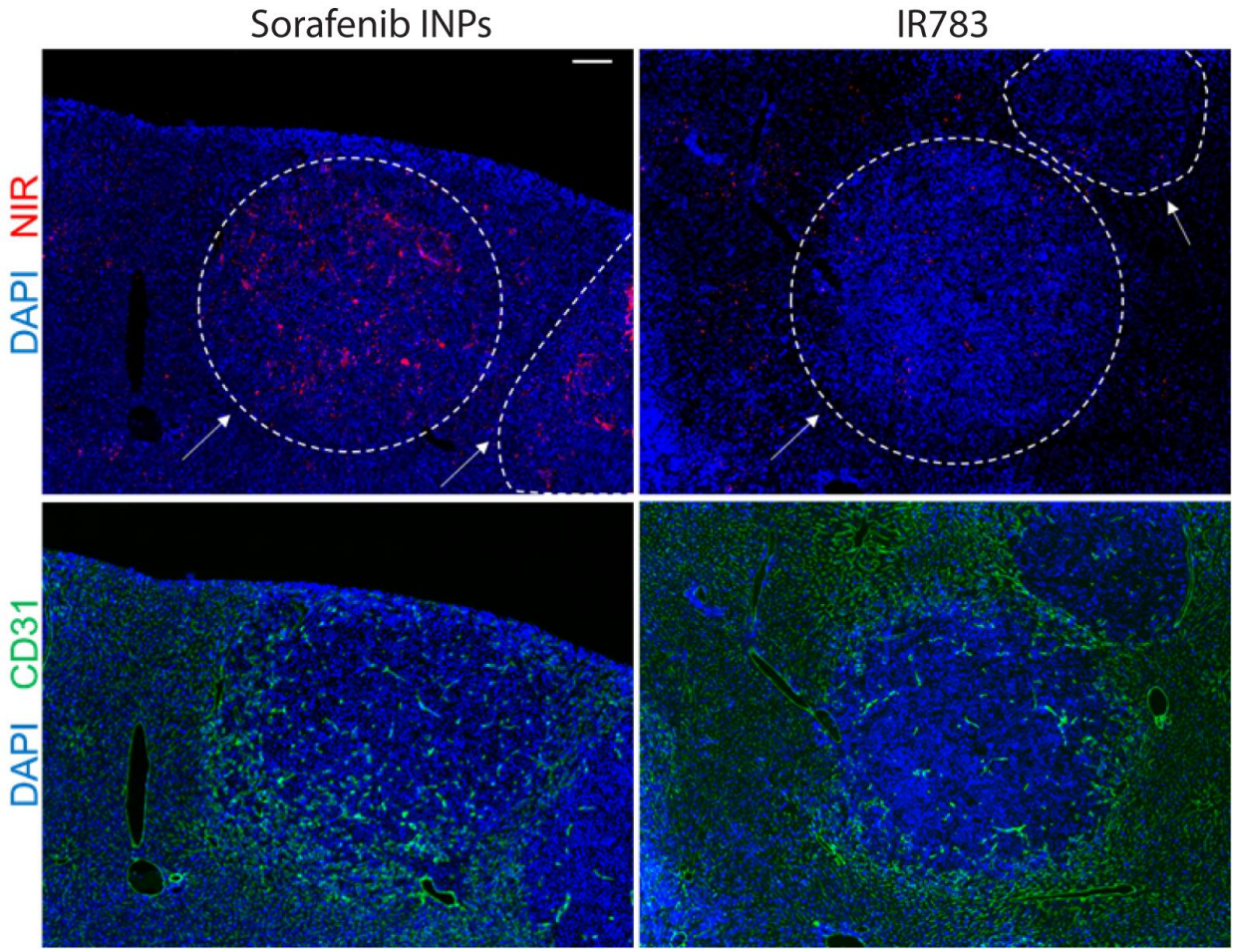
Supplementary Figure 18. Kinetics and efficiency of INP uptake in SK-136 cells. (a) Uptake of sorafenib INPs in SK-136 and MCF7 cells upon incubation with increasing concentrations of the nanoparticles for 2 h before imaging. Cells were stained DAPI. The percentage of NIR fluorescent cells out of all DAPI-stained cells was calculated using ImageJ software. **(b)** Kinetics of sorafenib INP uptake by SK-136 and MCF7 cells. A fixed concentration of 40 µg/ml of nanoparticles was added to the cells at time 0. Cells were imaged at the specified timepoint.



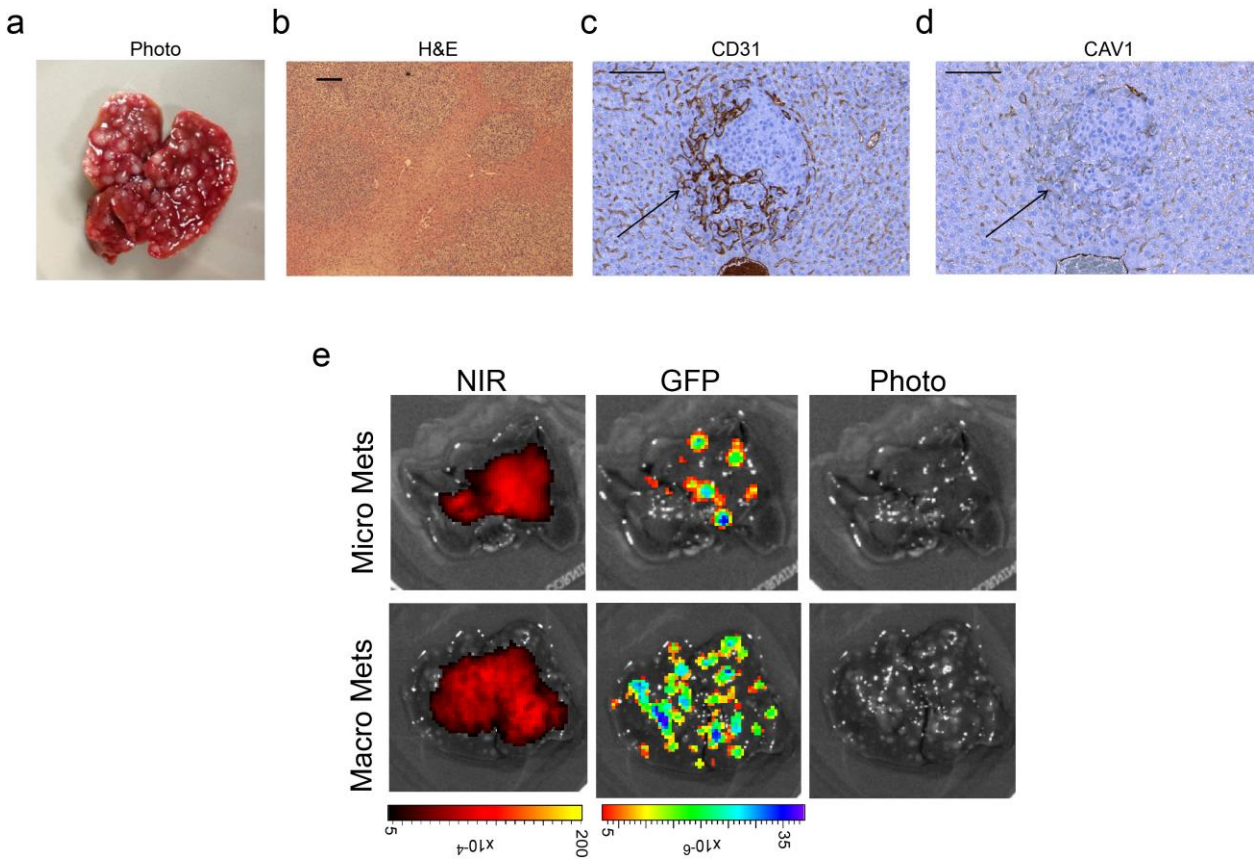
Supplementary Figure 19. **Biodistribution of indocyanine nanoparticles. (a)** Fluorescence images in vivo and organs ex vivo, 20 min after i.v. administration of sorafenib INPs, measured by IVIS. **(b)** Fluorescence images of organs ex vivo 24 h after i.v. administration of 3 different INPs. **(c)** Biodistribution of INPs quantified from the ex vivo fluorescence images in (b) as total fluorescence efficiency normalized by organ weight. **(d)** Blood plasma pharmacokinetics of i.v.-injected sorafenib solubilized in Kolliphor EL (SFB-IV) vs. sorafenib INPs (SFB INPs) in healthy mice, as measured by UV-VIS HPLC.



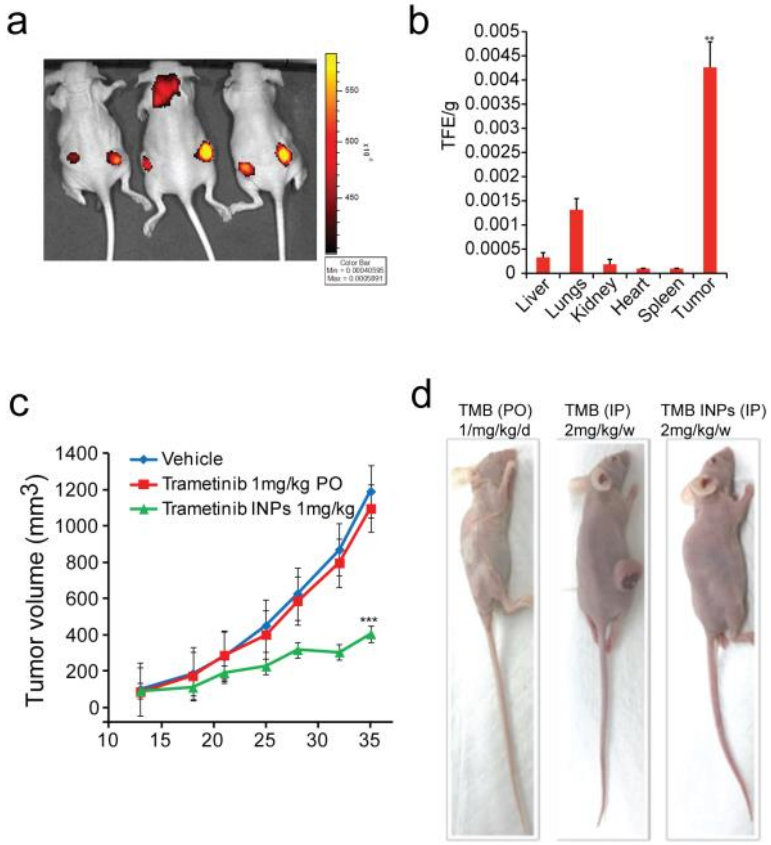
Supplementary Figure 20. **Sorafenib INP targeting in autochthonous liver cancer model. (a)** Color photograph of resected liver 21 days after inoculation. **(b)** Fluorescence image of GFP channel emission. (**c**) Hematoxylin and eosin stain. **(d)** Immunohistochemical stain for CD31. **(e)** Immunohistochemical stain for CAV1. All scale bars = 200 μm. **(f)** (Left) Accumulation of indocyanine (IR783) and INPs in livers of the genetically modified mouse model (top three livers) vs. normal livers (bottom two livers). (Right) GFP fluorescence images.



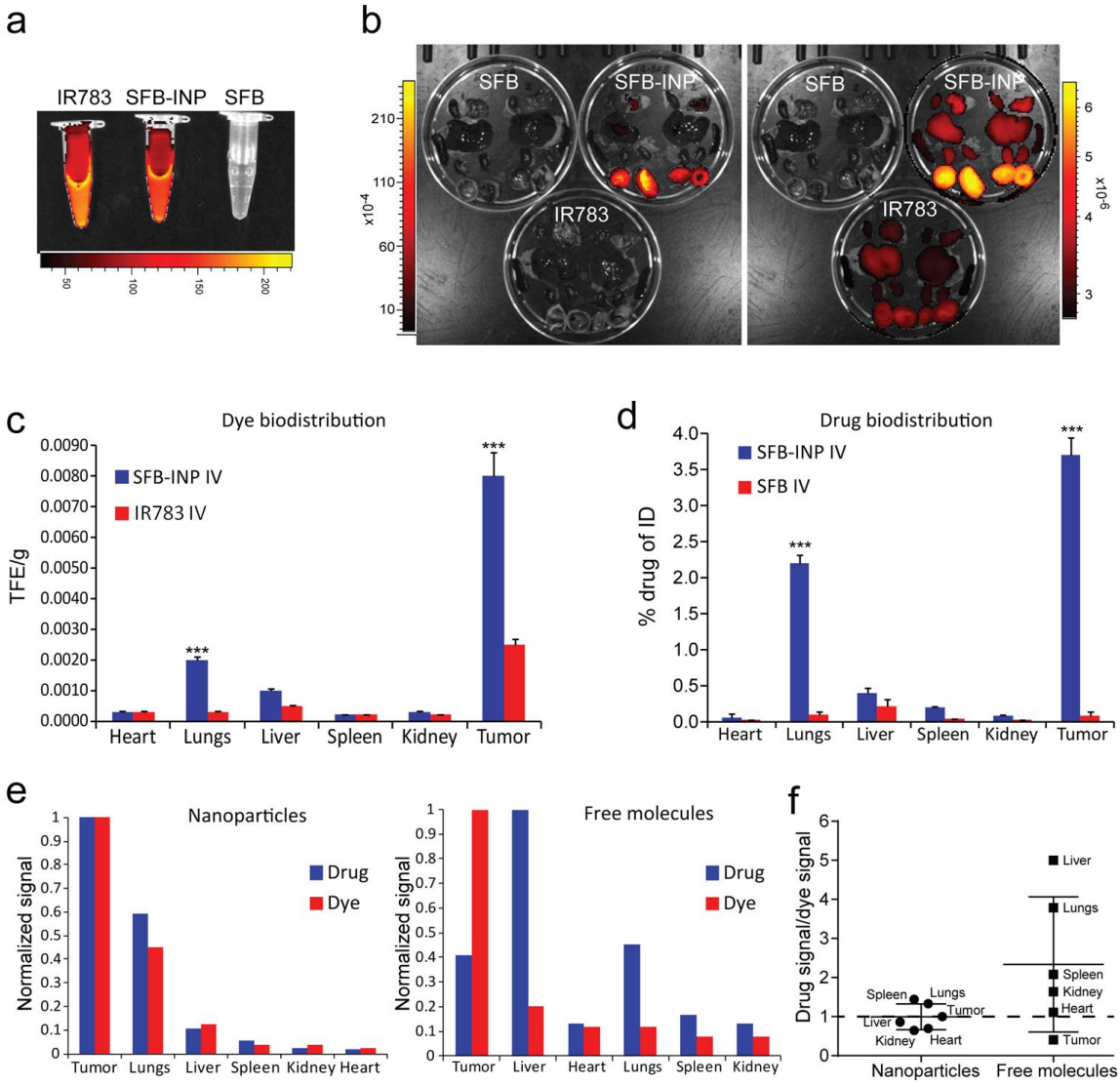
Supplementary Figure 21. Microdistribution of nanoparticles in autochthonous liver cancer model. Images of frozen tissue slices from the autochthonous liver cancer model 24 h after injections of sorafenib INPs or free IR783 dye. Red = NIR fluorescence, green = CD31 antibody for blood vessels, blue = DAPI for nuclear staining. Scale bar = 100 µm.



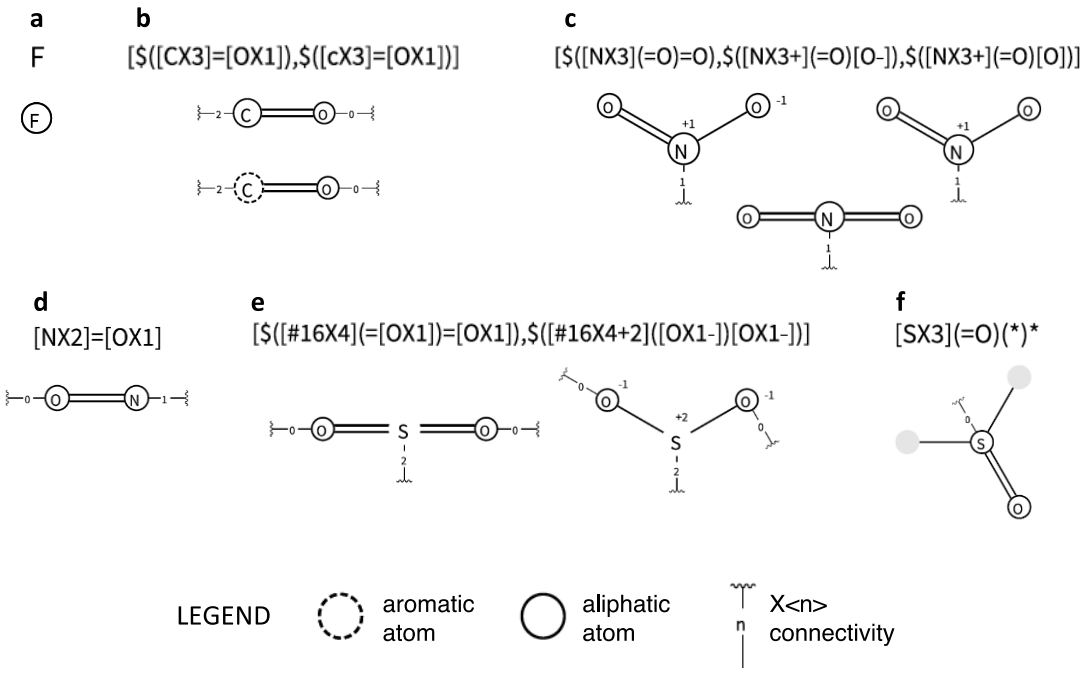
Supplementary Figure 22. Uveal melanoma liver metastasis model. (a) Photograph taken at 4 weeks after inoculation. **(b)** Hematoxylin and eosin stain of tumor tissue. **(c)** Immunohistochemical stain for CD31 at 3 weeks after inoculation. **(d)** Immunohistochemical stain for CAV1, 2 weeks after inoculation. Arrow indicates the tumor margin. All scale bars = 150 µm. **(e)** Images of livers from the uveal melanoma model 24 h after administration of nanoparticles to tumors (Top) 2 weeks after inoculation and (Bottom) 4 weeks after inoculation. (Left) Near-infrared channel. (Center) GFP fluorescence channel. (Right) Brightfield image.



Supplementary Figure 23. Trametinib INPs in HCT116 colon cancer model. (a) Near-infrared fluorescence of HCT116 mouse model imaged in vivo 24 h after i.v. administration of trametinib INPs. **(b)** Biodistribution of trametinib INPs 24 h after i.v. administration, calculated from ex vivo fluorescence images as total fluorescence efficiency divided by organ weight. **(c)** Tumor growth inhibition in response to i.v. injected nanoparticles given weekly, or free drug given orally weekly at equivalent doses (n=6). **(d)** Photographs of mice treated with trametinib administered orally daily, injected intraperitoneally weekly in DMSO, or injected intraperitoneally as INPs (TMB INPs) weekly, imaged 25 days after beginning treatments.



Supplementary Figure 24. **Drug and indocyanine biodistribution in HCT116 xenografts. (a)** Fluorescence images of sorafenib (SFB), sorafenib INPs (SFB-INP), and IR783 before injection. **(b)** Fluorescence images of tumors and organs ex-vivo 24 h after administration of INPs or SFB and IR783 in HCT116 xenografts. The two images are of the same organs with two different intensity thresholds applied. **(c)** Biodistribution of IR783 and INPs from ex vivo fluorescence images as total fluorescence efficiency normalized by organ weight. **(d)** Organ biodistribution of sorafenib 24h after i.v. injection of sorafenib INPs or free sorafenib, measured using HPLC. **(e)** Analysis of drug and dye co-localization after either injection of nanoparticles or free drug and dye (free molecules), from the organ distribution data in (c) and (d), normalized to the highest signal organ. **(f)** Evaluation of colocalization of drug and dye calculated by normalizing the drug signal to the dye signal for each tissue. Dashed line denotes 1:1 colocalization.



Supplementary Figure 25. Visual depiction of the matching substructures to SMARTS strings used in NHISS calculation. The calculation of NHISS descriptor involves substructure search of chemical groups with high intrinsic state values: SMARTS strings of (a) fluorine, (b) carbonyl, (c) nitro, (d) nitroso, (e) sulfonyl, and (f) sulfinyl functional groups. The NHISS descriptor is defined as the linear combination of the number of matches of the SMARTS-based substructure search. Substructure matches of fluorine, carbonyl, sulfinyl, and nitroso groups increment the NHISS value by 1; nitro and sulfonyl groups increment the NHISS value by 2. SMARTS editor and SMARTSviewer[44] were used to generate SMARTS strings and the images in this figure. For Python scripts for calculation of NHISS, refer to nano-drugbank repository (https://github.com/choderalab/nano-drugbank).

Supplementary Tables

Group	NAME	SpMAX4_ Bh(s)	ALOGP2	cLogP	Forms	NAME	SpMAX4_ Bh(s)	ALOGP2	cLogP	Forms
Training	Rapamycin	7.4	37.1	7.0	Yes	Silvesterol	5.0	5.7	3.4	No
	Docetaxel	7.4	7.1	4.1	Yes	Idelalisib	4.8	11.9	3.5	No
	Enzalutamide	8.0	18.8	3.4	Yes	Selumetinib	5.0	10.5	3.7	No
	Fulvestrant	8.0	70.8	7.4	Yes	Erlotinib	4.7	18.6	4.2	No
	Ventetoclax	7.0	60.9	10.3	Yes	Camptotecin	5.0	3.0	0.6	No
	Paclitaxel	7.4	9.8	4.7	Yes	Sunitinib	4.8	9.0	3.0	No
	Sorafenib	7.4	17.4	5.5	Yes	Taselisib	4.9	6.1	2.0	No
	Trametinib	7.4	10.1	4.8	Yes	Lapatinib	5.1	39.1	5.8	No
Validation	Avagacestat	8.0	16.0	3.1	Yes	Forskolin	6.4	0.5	1.3	No
	Dutasteride	8.0	32.5	4.9	Yes	Talazoparib	5.1	6.1	0.4	No
	Regorafenib	8.0	19.2	5.2	Yes	Ipatasertib	4.9	10.3	2.9	No
	RO4929097	8.0	10.8	3.6	Yes	Luminespib	5.2	16.1	2.8	No
	TAK-632	8.0	32.9	5.5	Yes	Apitolisib	5.0	1.8	1.3	No
	Celecoxib	7.7	20.7	4.4	Yes	Biminetinib	6.5	7.7	3.2	No
	Dabrafenib	7.7	34.2	4.6	Yes	Pazopanib	5.0	14.9	3.5	No
	Valrubicin	7.4	9.3	3.4	Yes	Dactolisib	4.9	28.8	5.8	No
	Pluripotin	7.4	22.9	4.3	Yes	ZSTK474	4.9	11.3	1.1	No
	Nilotinib	7.4	25.8	5.7	Yes	MGCD-265	4.9	31.0	5.1	No
	Vemurafenib	7.4	27.9	4.2	Yes	Valdecoxib	4.8	7.4	1.8	No
	Everolimus	7.4	35.5	7.4	Yes	Panobinostat	4.8	8.0	2.6	No
	Tanespimycin	7.4	5.0	3.0	Yes	Gefitinib	4.8	17.7	5.5	No
	BMS-777607	7.4	11.0	3.7	Yes	INK128	4.8	3.1	1.4	No
	Glimepiride	7.4	14.3	4.0	Yes	AZD-4547	4.7	17.7	4.4	No
	Tacrolimus	7.4	21.7	5.8	Yes	Voxtalisib	4.6	0.7	0.7	No
	ABT737	7.0	67.2	10.0	Yes	Verapamil	4.5	30.6	4.5	No
	Glibenclamide	7.0	17.1	4.2	Yes	Terbinafine	4.3	28.5	6.0	No
	Avasimibe	4.8	67.7	9.7	Yes	Ospemifene	4.7	31.6	5.6	No
						Bithionol	4.9	36.9	6.2	No
						Probucol	4.6	94.6	11.0	No
						Cholesterol	4.2	54.4	9.5	No
						Mubritinib	5.2	36.7	5.4	No
						Pyrene *	4.1	22.2	4.9	No
						Tetraphenyl ethylene *	4.68	45.82	7.3	No
						Tazarotene	4.5	24.5	6.1	No
						Cholecalciferol	4.2	60.1	9.47	No

*** Not a drug, but shown here for chemical diversity.**

Supplementary Table 1. **Drugs experimentally tested for nanoparticle formation with indocyanine.** SpMAX4_Bh(s) = the 4th largest eigenvalue of the Burden matrix weighted by the intrinsic state. ALOGP2 = a chemical descriptor for hydrophobicity. Group = whether each drug belongs to the training or validation set for QSAP Model 1 analysis for INP formation. Calculations of SpMAX4_Bh(s) and ALOGP2 were performed with Dragon 6. cLogP values were calculated with Chem3D.

Group	Intrinsic state
-F	8.0
=O	7.0
-OH	6.0
≡N	6.0
=N	5.0
-Cl	4.1
=S	3.7
-O-	3.5
-SH	3.2
=CH2	3.0
=N-	3.0
-Br	2.8
=C=	2.5
-I	2.1
>N-	2.0
-S-	1.8
>C=	1.7
>CH2	1.5
>CH-	1.3
>C<	1.3

Supplementary Table 2. Intrinsic state values of chemical groups. The intrinsic state of chemical groups calculated using the equation: $I_i = \frac{(2/L_i)^2 \times \delta_i^v + 1}{\delta_i}$ where L is the principal quantum number of the atom, δ^v is the number of valence electrons, and δ is the number of sigma electrons of the atom.

	NHISS ≥ 4	**NHISS < 4**		
SpMAX4_Bh(s) ≥ 7.00	True Positive 145	False Negative 2	**TPR (Sensitivity)** 0.986	**FNR** 0.014
SpMAX4_Bh(s) < 7.00	False Positive 3	True Negative 280	**FPR** 0.011	**TNR (Specificity)** 0.989
	FDR 0.020	**NPV** 0.993	**Accuracy** 0.988	
	PPV (Precision) 0.980	**FOR** 0.007		

Supplementary Table 3. Confusion matrix for self-assembly categorization of 430 compounds with NHISS and SpMAX4_Bh(s) descriptors. This matrix assesses the degree to which the NHISS descriptor reflects the SpMAX4_Bh descriptor. TPR = true positive rate, which is the sensitivity of the method, FPR = false positive rate, FNR = false negative rate, TNR = true negative rate, which is the specificity of the method, FDR = false discovery rate, PPV = positive predictive value, which is the precision of the method, NPV = negative predictive value, FOR = false omission rate.

		Confusion Matrix				
		TP	TN	FP	FN	Accuracy
Experiment[1] vs. SpMAX4_Bh(s)[2]	N = 16 (training set)	8	8	0	0	1.00
	N = 36 (validation set)	18	18	0	0	1.00
	N = 60 (all experimental data)	26	33	0	1	0.98
Experiment[1] vs. NHISS[3]	N = 52 (training set)	26	26	0	1	0.98
	N = 8 (validation set)	0	7	0	1	0.88
	N = 60 (all experimental data)	26	33	0	2	0.97

1 Binary categorization of INP-forming and not INP-forming drugs based on experimental observation.
[2] Binary categorization of drugs with SpMAX4_Bh(s) ≥ 7.0 (predicted to form INPs) and SpMAX4_Bh(s) < 7.0 (predicted to not form INPs).
[3] Binary categorization of drugs with NHISS ≥ 4 (predicted to form INPs) and NHISS < 4 (predicted not to form INPs).

Supplementary Table 4. Confusion matrix analysis of SpMAX4_Bh(s) and NHISS descriptors of QSAP model 1, compared to experimental results of nanoparticle formation with IR783. Cut-off value of SpMAX4_Bh(s) was selected based on experimental training set (N=16) and cut-off value for NHISS was selected to match the classification of SpMAX4_Bh(s), based on available experimental data at the time (N=52) and eight more drugs were tested later that consititute the validation set for classification by NHISS.

NAME	GETAWAY R4e	Observed Size (nm)	Predicted size (nm)	Deviation (nm)	Group
Paclitaxel	2.00	86	95.37	-9.37	Training
Docetaxel	2.52	112	120.48	-8.48	Training
Enzalutamide	2.00	115	95.51	19.49	Training
Vemurafenib	1.44	70	68.66	1.34	Training
Trametinib	1.35	55	64.26	-9.26	Training
Sorafenib	1.58	80	75.08	4.92	Training
Fulvestrant	2.46	120	118.02	1.98	Training
Nilotinib	1.48	70	70.59	-0.59	Training
ABT737	1.87	75	89.38	-14.38	Validation
Avagacestat	1.61	85	76.92	8.08	Validation
BMS-777607	1.43	77	68.08	8.92	Validation
Celecoxib	1.38	75	65.71	9.29	Validation
Rapamycin	2.39	129	114.21	14.79	Validation
Dabrafenib	1.82	95	86.72	8.28	Validation
Dutasteride	2.63	124	125.99	-1.99	Validation
Everolimus	2.21	125	105.75	19.25	Validation
Glimepiride	2.23	136	106.86	29.14	Validation
Glybenclamide	1.81	109	86.63	22.37	Validation
Venetoclax	2.07	96	99.09	-3.09	Validation
Regorafenib	1.65	82	78.71	3.30	Validation
RO4929097	1.84	100	87.79	12.21	Validation
Tacrolimus	2.31	118	110.58	7.42	Validation
TAK-632	1.64	82	77.98	4.02	Validation
Tanespimycin	1.88	81	89.96	-8.96	Validation
Valrubicin	1.90	87	90.68	-3.68	Validation
Pluripotin	1.73	90	82.57	7.43	Validation

Supplementary Table 5. Drugs experimentally tested in particle size study. Descriptor values, observed diameters, and size prediction results according to QSNAP Model 2 are given in the table. Group = whether the drug was placed into the training or validation set for QSNAP Model 2 analysis.

Group 1	Group 2	Group 3	Group 4	Group 5
Doxorubicin	Lapatinib	Dabrafenib	Ponatinib	Ventoclax
Pemetrexed	Gefitinib	Glibenclamide	ABT737	Docetaxel
Methotrexat	Valdecoxib	Glimepride	Cobimetinib	Trametinib
Gemcitabine	ZSTK474	Vemurafenib	Apprepitant	Sorafenib
Irinotecan	Binimetinib			Rapamycin
Vincristine	Luminespib			Paclitaxel
Vinblastine	Pazopanib			Enzalutamide
	Forskolin			Fulvestrant
	Apitolisib			ABT737
	Ipatasertib			Avagacestat
	Panobinostat			BMS-777607
	Voxtalisib			Celecoxib
	Talazoparib			Dutasteride
	INK128			Everolimus
	AZD-4547			Nilotinib
	Dactolisib			Pluripotin
	Terbinafine			Regorafenib
	MGCD-265			RO4929097
	Silvestrol			Tacrolimus
	Idelalisib			TAK-632
	Selumetinib			Tanespimycin
	Taselisib			Valrubicin
	Erlotinib			Avasimib
	Camptothecin			
	Sunitinib			
	Verapamil			

Table S6. **Experimentally tested compounds and corresponding decision tree groups**. Group 1 = water soluble amphiphilic drugs. Group 2 = water insoluble drugs with NHISS values below 4. Group 3 = water insoluble drugs with NHISS values above 4 and acidic pKa values. Group 4 = water insoluble drugs with NHISS values above 4 and basic pK_a values. Group 5 = water insoluble drugs with NHISS values above 4 and weak basic pK_a values. Green = forms nanoparticles. Red = does not form nanoparticles. Yellow = forms nanoparticles that are generally unstable at pH 7.